Novelty and Diversity in Recommender Systems

Pablo Castells, Neil Hurley, and Saúl Vargas

1 Introduction

Accurately predicting users' interests was the main direct or implicit drive of the recommender systems field in roughly the first decade and a half of the field's development. A wider perspective towards recommendation utility, including but beyond prediction accuracy, started to appear in the literature by the beginning of the 2000s [1, 2], taking views that began to realize the importance of novelty and diversity, among other properties, in the added value of recommendation [3, 4]. This realization grew progressively, reaching an upswing of activity by the turn of the past decade [5–9]. Today we might say that novelty and diversity have become an increasingly frequent part of evaluation practice. They are included increasingly often among the reported effectiveness metrics of new recommendation approaches, and are explicitly targeted by algorithmic innovations time and again [10–12]. It seems difficult to conceive of progress in the recommender systems field without considering these dimensions and further developing our understanding of them. Even though dealing with novelty and diversity remains an active area of research

This chapter results from work performed at UAM and is not associated with Amazon.

P. Castells (✉)
Universidad Autónoma de Madrid and Amazon, Madrid, Spain
e-mail: pablo.castells@uam.es

N. Hurley
University College Dublin, Dublin, Ireland
e-mail: neil.hurley@ucd.ie

S. Vargas
Infogrid, Tallinn, Estonia
e-mail: hello@saulvargas.es

F. Ricci et al. (eds.), *Recommender Systems Handbook*,
https://doi.org/10.1007/978-1-0716-2197-4_16

and development, considerable progress has been achieved in these years in terms of the development of enhancement techniques, evaluation metrics, methodologies, and theory.

In this chapter we analyze the different motivations, notions and perspectives under which novelty and diversity can be understood and defined (Sect. 2). We revise the evaluation procedures and metrics which have been developed in this area (Sect. 3), as well as the algorithms and solutions to enhance novelty and/or diversity (Sect. 4). We analyze the relationship with the prolific stream of work on diversity in Information Retrieval, as a confluent area with recommender systems, and discuss a unifying framework that aims to provide a common basis as comprehensive as possible to explain and interrelate different novelty and diversity perspectives (Sect. 5). We discuss further connections to bias and fairness in Sect. 6. We show some empirical results that illustrate the behavior of metrics and algorithms (Sect. 7), and close the chapter with a summary and discussion of the progress and perspectives in this area, and directions for future research (Sect. 8).

2 Novelty and Diversity in Recommender Systems

Novelty can be generally understood as the difference between present and past experience, whereas diversity relates to the internal differences within parts of an experience. The difference between the two concepts is subtle and close connections can in fact be established, depending on the point of view one may take, as we shall discuss. The general notions of novelty and diversity can be particularized in different ways. For instance, if a music streaming service recommends us a song we have never heard before, we would say this recommendation brings some novelty. Yet if the song is, say, a very canonical music type by some very well known singer, the involved novelty is considerably less than we would get if the author and style of the music were also original for us. We might also consider that the song is even more novel if, for instance, few of our friends know about it. On the other hand, a music recommendation is diverse if it includes songs of different styles rather than different songs of very similar styles, regardless of whether the songs are original or not for us. Novelty and diversity are thus to some extent complementary dimensions, though we shall seek and discuss in this chapter the relationships between them.

The motivations for enhancing the novelty and diversity of recommendations are manifold, as are the different angles one may take when seeking these qualities. This is also the case in other fields outside information systems, where novelty and diversity are recurrent topics as well, and considerable efforts have been devoted to casting clear definitions, equivalences and distinctions. We therefore start this chapter by overviewing the reasons for and the possible meanings of novelty and diversity in recommender systems, with a brief glance at related perspectives in other disciplines.

2.1 Why Novelty and Diversity in Recommendation

Bringing novelty and diversity into play as target properties of the desired outcome means taking a wider perspective on the recommendation problem concerned with final actual recommendation utility, rather than a single quality such as accuracy [3]. Novelty and diversity are not the only dimensions of recommendation utility one should consider aside from relevance (see e.g. Chap. 15 for a comprehensive survey), but they are fundamental ones [13]. The motivations for enhancing novelty and diversity in recommendations are themselves diverse, and can be founded in the system, user and business perspectives.

2.1.1 System Perspective

From the system point of view, user actions as implicit evidence of user needs involve a great deal of uncertainty as to what the actual user preferences really are. User clicks and purchases are certainly driven by user interests, but identifying what exactly in an item attracted the user, and generalizing to other items, involves considerable ambiguity. On top of that, system observations are a very limited sample of user activity, whereby recommendation algorithms operate on significantly incomplete knowledge. Furthermore, user interests are complex, highly dynamic, context-dependent, heterogeneous and even contradictory. Predicting the user needs is therefore an inherently difficult task, unavoidably subject to a non-negligible error rate.

Diversity can be a good strategy to cope with this uncertainty and optimize the chances that at least some item pleases the user, by widening the range of possible item types and characteristics at which recommendations aim, rather than bet for a too narrow and risky interpretation of user actions. For instance, a user who has rated the movie “Rango” with the highest value may like it because—in addition to more specific virtues—it is a cartoon, a western, or because it is a comedy. Given the uncertainty about which of the three characteristics may account for the user preference, recommending a movie of each genre generally pays off more than recommending, say three cartoons. Three hits do not necessarily bring three times the gain of one hit—e.g. the user might rent just one recommended movie anyway—, whereas the loss involved in zero hits is considerably worse than achieving a single hit. From this viewpoint we might say that diversity is not necessarily an opposing goal to accuracy, but in fact a strategy to optimize the gain drawn from accuracy and relevance in matching true user needs in an uncertain environment.

2.1.2 User Perspective

From the user perspective, novelty and diversity are generally desirable per se, as a direct source of user satisfaction [4, 14–17]. Consumer behaviorists have long studied the natural variety-seeking drive in human behavior [18]. The explanation of this drive is commonly divided into direct and derived motivations. The former refer to the inherent satisfaction obtained from "novelty, unexpectedness, change and complexity" [19], and a genuine "desire for the unfamiliar, for alternation among the familiar, and for information" [20], linking to the existence of an ideal level of stimulation, dependent on the individual. Satiation and decreased satisfaction results from the repeated consumption of a product or product characteristic in a decreasing marginal value pattern [21]. As preferences towards discovered products are developed, consumer behavior converges towards a balance between alternating choices and favoring preferred products [22]. Derived motivations include the existence of multiple needs in people, multiple situations, or changes in people's tastes [18].

Some authors also explain diversity-seeking as a strategy to cope with the uncertainty about one's own future preference when one will actually consume the choices [23], as e.g. when we choose books and music for a trip. Moreover, novel and diverse recommendations enrich the user experience over time, helping expand the user's horizon. It is in fact often the case that we approach a recommender system with the explicit intent of discovering something new, developing new interests, and learning. The potential problems of the lack of diversity which may result from too much personalization has likewise raised issues in recent years; the concern for so-called echo chambers and filter bubbles [24] is familiar nowadays to the general public. Reconciling personalization with a healthy degree of diversity is certainly part of any approach to deal with these problems.

2.1.3 Business Perspective

Diversity and novelty also find motivation in the underlying businesses in which recommendation technologies are deployed. Customer satisfaction indirectly benefits the business in the form of increased activity, revenues, and customer loyalty. Beyond this, product diversification is a well-known strategy to mitigate risk and expand businesses [25]. Moreover, selling in the long tail is a strategy to draw profit from market niches by selling less of more and getting higher profit margins on cheaper products [26].

2.1.4 The Limits of Novelty and Diversity

All the above general considerations can be of course superseded by particular characteristics of the specific domain, the situation, and the goal of the recommendations, for some of which novelty and diversity are indeed not always needed. For instance, getting a list of similar products to one we are currently inspecting (e.g. a TV set, a holiday rental, etc.) may help us refine our choice among a large set of very similar options. Recommendations can serve as a navigational aid in this type of situation. In other domains, it makes sense to consume the same or very similar items again and again, such as grocery shopping, clothes, etc. The added value of recommendation is probably more limited in such scenarios though, where other kinds of tools may solve our needs (catalog browsers, shopping list assistants, search engines, etc.), and even in these cases we may appreciate some degree of variation in the mix every now and then. We briefly discuss some specific studies on the motivation and effect of recommendation novelty and diversity on actual users later in Sect. 4.7.

2.2 Defining Novelty and Diversity

Novelty and diversity are different though related notions, and one finds a rich variety of angles and perspectives on these concepts in the recommender system literature, as well as other fields such as sociology, economics, or ecology. As pointed out at the beginning of this section, novelty generally refers, broadly, to the difference between present and past experience, whereas diversity relates to the internal differences within parts of an experience. Diversity generally applies to a set of items or "pieces", and has to do with how different the items or pieces are with respect to each other. Variants have been defined by considering different pieces and sets of items. In the basic case, diversity is assessed in the set of items recommended to each user separately (and typically averaged over all users afterwards) [4]. But global diversity across sets of sets of items has also been considered, such as the recommendations delivered to all users [6, 27, 28], recommendations by different systems to the same user [29], or recommendations to a user by the same system over time [30].

The novelty of a set of items can be generally defined as a set function (average, minimum, maximum) on the novelty of the items it contains. We may therefore consider novelty as primarily a property of individual items. The novelty of a piece of information generally refers to how different it is with respect to "what has been previously seen" or experienced. This is related to diversity in that when a set is diverse, each item is "novel" with respect to the rest of the set. Moreover, a system that promotes novel results tends to generate global diversity over time in the user experience; and also enhances the global "diversity of sales" from the system perspective. Multiple variants of novelty arise by considering the fact that novelty is relative to a context of experience, as we shall discuss.

Different nuances have been considered in the concept of novelty. A simple definition of novelty can consist of the (binary) absence of an item in the context of reference (prior experience). We may use adjectives such as unknown or unseen for this notion of identity-based novelty [9]. Long tail notions of novelty are elaborations of this concept, as they are defined in terms of the number of users who would specifically know an item [7, 28, 31]. But we may also consider how different or similar an unseen item is with respect to known items, generally—but not necessarily—on a graded scale. Adjectives such as unexpected, surprising and unfamiliar have been used to refer to this variant of novelty [5, 11, 12]. Unfamiliarity and identitary novelty can be related by trivially defining similarity as equality, i.e. two items are "similar" if and only if they are the same item. Finally, the notion of serendipity is used to mean novelty plus a positive emotional response—in other words, an item is serendipitous if it is novel—unknown or unfamiliar—and relevant [32–34].

The present chapter is concerned with the diversity and novelty involved in recommendations, but one might also study the diversity (in tastes, behavior, demographics, etc.) of the end-user population, or the product stock, the sellers, or in general the environment in which recommenders operate. While some works in the field have addressed the diversity in user behavior [35, 36], we will mostly focus on those aspects a recommender system has a direct hold on, namely the properties of its own output.

2.3 Diversity in Other Fields

Diversity is a recurrent theme in several fields, such as sociology, psychology, economics, ecology, genetics or telecommunications. One can establish connections and analogies from some—though not all—of them to recommender systems, and some equivalences in certain metrics, as we will discuss.

Diversity is a common keyword in sociology referring to cultural, ethnic or demographic diversity [37]. Analogies to recommender system settings would apply to the user population, which is mainly a given to the system, and therefore not within our main focus here. In economics, diversity is extensively studied in relation to different issues such as the players in a market (diversity vs. oligopolies), the number of different industries in which a firm operates, the variety of products commercialized by a firm, or investment diversity as a means to mitigate the risk involved in the volatility of investment value [25]. Of all such concepts, product and portfolio diversity most closely relate to recommendation, as mentioned in Sect. 2.1.3, as a general risk-mitigating principle and/or business growth strategy.

Behaviorist psychology has also paid extensive attention to the human drive for novelty and diversity [18]. Such studies, especially the ones focusing on consumer behavior, provide formal support to the intuition that recommender system users may prefer to find some degree of variety and surprise in the recommendations they receive, as discussed in Sect. 2.1.2.

An extensive strand or literature is devoted to diversity in ecology as well, where researchers have worked to considerable depth on formalizing the problem, defining and comparing a wide array of diversity metrics, such as the number of species (richness), Gini-Simpson and related indices, or entropy [38, 39]. Such developments connect to aggregate recommendation diversity perspectives that deal with sets of recommendations as a whole, as we shall discuss in Sects. 3.5 and 5.3.3.

Finally, the issue of diversity has also attracted a great deal of attention in the Information Retrieval (IR) field. A solid body of theory, metrics, evaluation methodologies and algorithms has been developed in this scope in the last decades [40–46], including a dedicated search diversity task in four consecutive TREC editions starting in 2009 [47]. Search and recommendation are different problems, but have much in common: both tasks are about ranking a set of items to maximize the satisfaction of a user need, which may or may not have been expressed explicitly. It has in fact been found that the diversity theories and techniques in IR and recommender systems can be connected [48, 49], as we will discuss in Sect. 5.4. Given these connections, and the significant developments on diversity in IR, we find it relevant to include an overview of this work here, as we will do in Sects. 3 (metrics) and 4 (algorithms).

3 Novelty and Diversity Evaluation

The definitions discussed in the previous sections can only get a full, precise and practical meaning when one has given a specific definition of the metrics and methodologies by which novelty and diversity are to be measured and evaluated. We review next the approaches and metrics that have been developed to assess novelty and diversity, after which we will turn to the methods and algorithms proposed in the field to enhance them.

3.1 Notation

As is common in the literature, we will use the symbols i and j to denote items, u and v for users, $\mathcal{I}$ and $\mathcal{U}$ for the set of all items and users respectively. By $\mathcal{I}_u$ and $\mathcal{U}_i$ we shall denote, respectively, the set of all items u has interacted with, and the set of users who have interacted with i. In general we shall take the case where the interaction consists of rating assignment (i.e. at most one time per user-item pair), except where the distinction between single and multiple interaction makes a relevant difference (namely Sect. 5.2.1). We denote ratings assigned by users to items as $r(u,i)$, and use the notation $r(u,i) = \emptyset$ to indicate missing ratings, as in [50]. We shall use R to denote a recommendation to some user, and R_u whenever we wish or need to explicitly indicate the target user u to whom R is delivered—in other words, R will be a shorthand for R_u. By default, the definition of a metric will be given on a single recommendation for a specific target user. For notational simplicity, we omit as understood that the metric should be averaged over all users. Certain global metrics (such as aggregate diversity, defined in Sect. 3.5) are the exception to this rule: they directly take in the recommendations to all users in their definition, and they therefore do not require averaging. In some cases where a metric is the average of a certain named function (e.g. IUF for inverse user frequency, SI for self-information) on the items it contains, we will compose the name of the metric by prepending an "M" for "mean" (e.g. MIUF, MSI) in order to distinguish it from the item-level function.

3.2 Average Intra-List Distance

Perhaps the most frequently considered diversity metric and the first to be proposed in the area is the so-called average intra-list distance—or intra-list diversity, ILD (e.g. [2, 4, 51]). The intra-list diversity of a set of recommended items is defined as the average pairwise distance of the items in the set:

$$\mathrm{ILD} = \frac{1}{|R|(|R|-1)} \sum_{i \in R} \sum_{j \in R} d(i, j)$$

The computation of ILD requires defining a distance measure $d(i, j)$, which is thus a configurable element of the metric. Given the profuse work on the development of similarity functions in the recommender systems field, it is common, handy and sensible to define the distance as the complement of well-understood similarity measures, but nothing prevents the consideration of other particular options. The distance between items is generally a function of item features [4], though the distance in terms of interaction patterns by users has also been considered sometimes [52].

The ILD scheme in the context of recommendation was first suggested, as far as we are aware of, by Smyth and McClave [2], and has been used in numerous subsequent works (e.g. [4, 9, 51, 52]). Some authors have defined this dimension by its equivalent complement intra-list similarity ILS [4], which has the same relation to ILD as the distance function has to similarity, e.g. $\mathrm{ILD} = 1 - \mathrm{ILS}$ if $d = 1 - sim$.

3.3 Global Long-Tail Novelty

The novelty of an item from a global perspective can be defined as the opposite of popularity: an item is novel if few people are aware it exists, i.e. the item is far in the long tail of the popularity distribution [7, 31]. Zhou et al. [28] modeled popularity as the probability that a random user would know the item. To get a decreasing function of popularity, the negative logarithm provides a nice analogy with the inverse document frequency (IDF) in the vector-space Information Retrieval model, with users in place of documents and items instead of words, which has been referred to as inverse user frequency (IUF) [53]. Based on the observed user-item interaction, this magnitude can be estimated as $\mathrm{IUF} = -\log_2 |\mathcal{U}_i|/|\mathcal{U}|$. Thus the novelty of a recommendation can be assessed as the average IUF of the recommended items:

$$\mathrm{MIUF} = -\frac{1}{|R|} \sum_{i \in R} \log_2 \frac{|\mathcal{U}_i|}{|\mathcal{U}|} \tag{1}$$

The IUF formula also has a resemblance to the self-information measure of Information Theory, only for that to be properly the case, the probability should add to 1 over the set of items, which is not the case here. We discuss that possibility in Sect. 5.2.1.

3.4 User-Specific Unexpectedness

Long-tail novelty translates to non-personalized measures for which the novelty of an item is seen as independent of the target user. It makes sense however to consider the specific experience of a user when assessing the novelty carried by an item that is recommended to her, since the degree to which an item is more or less familiar can greatly vary from one user to the next.

Two perspectives can be considered when comparing an item to prior user experience: the item identity (was this particular item seen before?) or the item characteristics (were the attributes of the item experienced before?). In the former view, novelty is a Boolean property of an item which occurs or not in its totality, whereas the latter allows to appreciate different degrees of novelty in an item even if it was never, itself, seen before.

It is not straightforward to define identity-based novelty on an individual user basis. In usual scenarios, if the system observes the user interact with an item, it will avoid recommending her this item again.[1] This is a rather trivial feature and does not need to be evaluated—if anything, just debugged (e.g. for near-duplicate detection). We may therefore take it for granted, except in particular scenarios where users recurrently consume items—where on the other hand a recommender system may have a more limited range for bringing added value. It would be meaningful though to assess the Boolean novelty of an item in probabilistic terms, considering the user activity outside the system, which in a detailed sense is of course impractical. Long tail novelty can be seen as a proxy for this notion: a user-independent estimate of the prior probability that the user—any user—has seen the item before. Finer, user-specific probability estimation approaches could be explored but have not, to the best of our knowledge, been developed in the literature so far.

An attribute-based perspective is an easier-to-compute alternative for a user-specific novelty definition. Taking the items the user has been observed to encounter, the novelty of an item can be defined in terms of how different it is to the previously encountered items, as assessed by some distance function on the item properties. This notion reflects how unfamiliar, unexpected and/or surprising an item may be based on the user's observed experience. The set-wise distance to the profile items can be defined by aggregation of the pairwise distances by an average, minimum, or other suitable function. For instance, as an average:

[1] Of course, what "interaction" means and to what extent it will inhibit future recommendations is application-dependent, e.g. an online store may recommend an item the user has inspected but not bought.

$$\text{Unexp} = \frac{1}{|R||\mathcal{I}_u|} \sum_{i \in R} \sum_{j \in \mathcal{I}_u} d(i, j) \,.$$

Some authors have generalized the notion of unexpectedness to the difference of a recommendation with respect to an expected set of items, not necessarily the ones in the target user profile, thus widening the perspective on what "expected" means [5, 12, 32, 54]. For instance, Murakami et al. [32] define the expected set as the items recommended by a "primitive" system which is supposed to produce unsurprising recommendations. The difference to the expected set can be defined in several ways, such as the ratio of unexpected recommended items:

$$\text{Unexp} = |R - EX|/|R| \tag{2}$$

EX being the set of expected items. Other measures between the recommended and expected set include the Jaccard distance, the centroid distance, the difference to an ideal distance, etc. [5].

3.5 Inter-Recommendation Diversity Metrics

In the early 2010s Adomavicius and Kwon [6, 27] proposed measuring the so-called aggregate diversity of a recommender system. This perspective is different from all the metrics described above in that it does not apply to a single set of recommended items, but to all the output a recommender system produces over a set of users. It is in fact a quite simple metric which counts the total number of items that the system recommends.

$$\text{Aggdiv} = \left| \bigcup_{u \in \mathcal{U}} R_u \right| \tag{3}$$

A version Aggdiv@k of the metric can be defined by taking R_u as the top k items recommended to u. Since it applies to the set of all recommendations, aggregate diversity does not need to be averaged over users, differently from most other metrics mentioned in these pages.

Aggregate diversity is a relevant measure to assess to what extent an item inventory is being exposed to users. The metric, or close variations thereof, have also been referred to as item coverage in other works [1, 29, 54–56] (see also Chap. 15). This concept can be also related to traditional diversity measures such as the Gini coefficient, the Gini-Simpson's index, or entropy [38, 39], which are commonly used to measure statistical dispersion in such fields as ecology (biodiversity in ecosystems), economics (wealth distribution inequality), or sociology (e.g. educational attainment across the population).

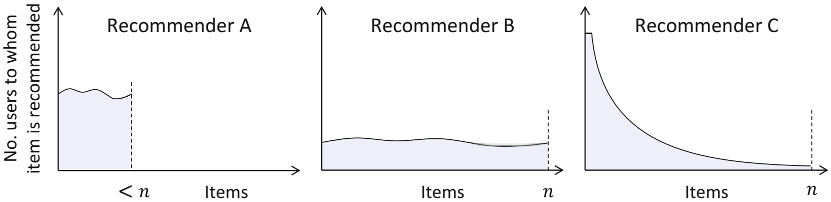


Fig. 1 By aggregate diversity, recommender B does better than A, but just as well as C. The Gini index, Simpson-Gini and Entropy notice however that B distributes recommendations more evenly over items than C

Mapped to recommendation diversity, such measures take into account not just whether items are recommended to someone, but to how many people and how even or unevenly distributed. To this extent they serve a similar purpose as aggregate diversity as measures of the concentration of recommendations over a few vs. many items. While aggregate diversity counts item exposure to users in a binary way, the Gini index, Simpson-Gini and entropy are more informative as they are sensitive to the amount of users that are recommended each item. Figure 1 illustrates this.

For instance, Fleder and Hosanagar [35] measure sales concentration by the Gini index, which Gunawardana and Shani (Chap. 15) formulate as:

$$\text{Gini} = \frac{1}{|\mathcal{I}| - 1} \sum_{k=1}^{|\mathcal{I}|} (2k - N - 1)\, p(i_k|s)$$

where $p(i_k|s)$ is the probability of the k-th least recommended item being drawn from the recommendation lists generated by a system s:

$$p(i|s) = \frac{|\{u \in \mathcal{U} \mid i \in R_u\}|}{\sum_{j \in \mathcal{I}} |\{u \in \mathcal{U} \mid j \in R_u\}|}$$

The Gini index and aggregate diversity have been used in subsequent work such as [57, 58]. Other authors (e.g. [36] or Chap. 15) suggest the Shannon entropy with similar purposes:

$$\text{H} = -\sum_{i \in \mathcal{I}} p(i|s) \log_2 p(i|s)$$

Related to this, Zhou et al. [28] observe the diversity of the recommendations across users. They define inter-user diversity (IUD) as the average pairwise Jaccard distance between recommendations to users. In a quite equivalent reformulation of this measure we may define the novelty of an item as the ratio of users to which it is not recommended:[2]

$$\text{IUD} = \frac{1}{|R|}\sum_{i \in R}\frac{|\{v \in \mathcal{U} \mid i \notin R_v\}|}{|\mathcal{U}| - 1} = \frac{1}{|\mathcal{U}| - 1}\sum_{v \in \mathcal{U}}|R - R_v|/|R| \tag{4}$$

Since $|R - R_v|/|R| = 1 - |R \cap R_v|/|R \cup R_v|$, it can be seen that the difference between this definition and the Jaccard-based formulation is basically that the latter has $|R|$ instead of $|R \cup R_v|$ in the denominator, but the above formulation is interesting because it connects to the Gini-Simpson index, as we will show in Sect. 5.3.3.

With a similar metric structure, Bellogín et al. [29] measure the inter-system diversity (ISD), i.e. how different the output of a system is with respect to other systems, in settings where several recommenders are operating. This can be defined as the ratio of systems that do not recommend each item:

$$\text{ISD} = \frac{1}{|R|}\sum_{i \in R}\frac{|\{s \in \mathcal{S} \mid i \notin R^s\}|}{|\mathcal{S}| - 1} = \frac{1}{|\mathcal{S}| - 1}\sum_{s \in \mathcal{S}}|R - R^s|/|R| \tag{5}$$

where $\mathcal{S}$ is the set of recommenders in consideration, and R^s denotes the recommendation to the target user by a system $s \in \mathcal{S}$. This metric thus assesses how different the output of a recommender system is with respect to alternative algorithms. This perspective can be useful, for instance, when an application seeks to distinguish itself from the competition, or when selecting an algorithm to add to an ensemble.

In a different angle, Lathia et al. [30] consider the time dimension in novelty and diversity. Specifically, they study the diversity between successive recommendations by a system to a user, as the ratio of items that were not recommended before:

$$\text{TD} = |R - R'|/|R| \tag{6}$$

The authors distinguish the difference between consecutive recommendations, and the difference between the last recommendation and all prior recommendations. In the former case (which they name "temporal diversity") R' is the recommendation immediately preceding R, and in the latter ("temporal novelty") R' is the union of all recommendations to the target user preceding R. In both cases, the metric gives a perspective of the ability of a recommender system to evolve with the changes in the environment in which it operates, rather than presenting users the same set of items over and over again.

[2] Note that we normalize IUD by $|\mathcal{U}| - 1$ because all items in R are recommended to at least one user (the target of R), therefore if we normalized by $|\mathcal{U}|$, the value of the metric for the optimal recommendation would be $(|\mathcal{U}| - 1)/|\mathcal{U}| < 1$. Put another way, $v \in \mathcal{U}$ in the numerator could be as well written as $v \in \mathcal{U} - \{u\}$, which would call for normalizing by $|\mathcal{U} - \{u\}| = |\mathcal{U}| - 1$. The difference is negligible in practice though, and we believe both forms of normalization would be acceptable. The same rationale applies to Eq. 5 next.

Note that IUD, ISD and TD fit as particular cases under the generalized unexpectedness scheme [5] described in the previous section (Eq. 2), where the set EX of expected items would be the items recommended to other users by the same system ($EX = R_v$), to the same user by other systems ($EX = R^s$), or to the same user by the same system in the past ($EX = R'$). One difference is that IUD and ISD take multiple sets EX for each target user (one per user v and one per system s respectively), whereby these metrics involve an additional average over such sets.

In a different perspective, Sanz-Cruzado and Castells [59] research diversity as a global notion across users in the context of social networks, and the effects that recommending people to befriend can have on the evolution of the network structure as a whole. Drawing upon concepts in social network analysis, they consider different diversity angles and how to enhance them. Focusing on weak links as a source of novelty, they analyze the effects that recommending them can have on the diversity of information flowing through the network.

3.6 Specific Methodologies

As an alternative to the definition of special-purpose metrics, some authors have evaluated the novelty or diversity of recommendations by accuracy metrics on a diversity-oriented experimental design. For instance, Hurley and Zhang [8] evaluate the diversity of a system by its ability to produce accurate recommendations of difficult items, "difficult" meaning unusual or infrequent for a user's typical observed habits. Specifically, a data splitting procedure is set up by which the test ratings are selected among a ratio of the top most different items rated by each user, "different" being measured as the average distance of the item to all other items in the user profile. The precision of recommendations in such a setting thus reflects the ability of the system to produce good recommendations made up of novel items. A similar idea is to select the test ratings among cold, non-popular long tail items. For instance, Zhou et al. [28] evaluate accuracy on the set of items with less than a given number of ratings. Shani and Gunawardana also discuss this idea in Chap. 15.

3.7 Diversity vs. Novelty vs. Serendipity

Even though the distinction between novelty and diversity is not always a fully clean-cut line, We may propose a classification of the metrics described so far as either novelty or diversity measures. ILD can be considered the genuine metric for diversity, the definition of which it applies to the letter. We would also class inter-recommendation metrics (Sect. 3.5) in the diversity type, since they assess how different are recommendations to each other. They do so at a level above an individual recommendation, by (directly or indirectly) comparing sets of recommended items rather than item pairs.

On the other hand, we may consider that long tail and unexpectedness fit in the general definition of novelty: unexpectedness explicitly measures how different each recommended item is with respect to what is expected, where the latter can be related to previous experience. And long tail non-popularity defines the probability that an item is different (is absent) from what a random user may have seen before. The methodologies discussed in the previous section can also be placed in the novelty category, as they assess the ability to properly recommend novel items.

It should also be noted that several authors target the specific concept of serendipity as the conjunction of novelty and relevance [8, 28, 32, 34, 54, 60]. In terms of evaluation metrics, this translates to adding the relevance condition in the computation of the metrics described in Sects. 3.3 and 3.4. In other words, taking the summations over $i \in R \wedge i$ relevant to u in place of just $i \in R$ turns a plain novelty metric (long tail or unexpectedness) into the corresponding serendipity metric.

3.8 Information Retrieval Diversity

Differently (at least apparently) from the recommender systems field, diversity in IR has been related to an issue of uncertainty in the user query. Considering that most queries contain some degree of ambiguity or incompleteness as an expression of user needs, diversity is posited as a strategy to cope with this uncertainty by answering as many interpretations of the query as early as possible in the search results ranking. The objective is thus redefined from returning as many relevant results as possible to maximizing the probability that all users (all query interpretations) will get at least some relevant result. This principle is derived from reconsidering the independence assumption on document relevance, whereby returning relevant documents for different query interpretations pays off more than the diminishing returns from additional relevant documents for the same interpretation. For instance a polysemic query such as “table” might be interpreted as furniture or a database concept. If a search engine returns results in only one of the senses, it will satisfy 100% the users who were intending this meaning, and 0% the rest of users. But combining instead a balanced mix of both intents, results will likely meet the needs of most users (getting them well more than half-satisfied), in a typical search where a few relevant results are sufficient to satisfy the user need.

IR diversity metrics have been defined under the assumption that an explicit space of possible query intents (also referred to as query aspects or subtopics) can be represented. In general, the aspects for evaluation should be provided manually, as has been done in the TREC diversity task, where a set of subtopics is provided for each query, along with per-subtopic relevance judgments [47].

Probably the earliest proposed metric was subtopic recall [46], which simply consists of the ratio of query subtopics covered in the search results:

$$\mathrm{S-recall} = \frac{|\{z \in \mathcal{Z} \mid d \in R \wedge d \text{ covers } z\}|}{|\{z \in \mathcal{Z} \mid z \text{ is a subtopic of } q\}|}$$

where $\mathcal{Z}$ is the set of all subtopics. Later on the TREC campaign popularized metrics such as ERR-IA [42] and α-nDCG [44], and a fair array of other metrics have been proposed as well. For instance, based on the original definition of ERR in [42], the intent-aware version ERR-IA is:

$$\mathrm{ERR-IA} = \sum_{z} p(z|q) \sum_{d_k \in R} p(rel|d_k, z) \prod_{j=1}^{k-1} (1 - p(rel|d_j, z)) \tag{7}$$

where $p(z|q)$ takes into account that not all aspects need to be equally probable for a query, weighting their contribution to the metric value accordingly. And $p(rel|d, z)$ is the probability that document d is relevant to the aspect z of the query, which can be estimated based on the relevance judgments. E.g. for graded relevance Chapelle [42] proposed $p(rel|d, z) = 2^{g(d,z)-1}/2^{g_{max}}$, where $g(d, z) \in [0, g_{max}]$ is the relevance grade of d for the aspect z of the query. It is also possible to consider simpler mappings, such as a linear map $g(d, z)/g_{max}$, depending on how the relevance grades are defined [9].

Novelty, as understood in recommender systems, has also been addressed in IR, though perhaps not to as much extent as diversity. It is mentioned, for instance, in [61] as the ratio of previously unseen documents in a search result. It is also studied at the level of document sentences, in terms of the non-redundant information that a sentence provides with respect to the rest of the document [62]. Even though the concept is essentially the same, to what extent one may establish connections between the sentence novelty techniques and methodologies, and item novelty in recommendation is not obvious, but might deserve future research.

3.9 *Proportional Diversity*

Closely related to aspect-based diversity in IR, recommendation diversity can be formulated with respect to the different subareas that user interests typically encompass, that we may want to cover in recommendations. An interesting idea in this line is to aim at a coverage of user interest aspects in recommendations that is not necessarily uniform but proportional to a specific desired distribution. A reasonable such distribution can be the proportion of each aspect observed in each user's prior interaction history with items. Vargas et al. [63] and Steck [64] explore different ideas in this direction. The intent-aware scheme [40] in IR, of which the ERR-IA metric [42] described in the previous section is an example, also incorporates the proportionality principle through the $p(z|q)$ term (see Eq. 7), representing the importance of aspect z for the information need q—which we can translate to the interests of a user u in a recommendation context. We further discuss later in Sect. 5.4.2 the direct adaptation of IR diversity to recommender systems.

4 Novelty and Diversity Enhancement Approaches

Methods to enhance the novelty and diversity of recommendations are reviewed in this section. It is noteworthy that research in this area has accelerated over the last number of years. The work can be categorized into methods that re-rank an initial list to enhance the diversity/novelty of the top items; methods based on clustering; hybrid or fusion methods; and methods that consider diversity in the context of learning to rank objectives.

4.1 Result Diversification/Re-ranking

One common approach to enhance the diversity of recommendation is the diversification or re-ranking of the results returned by an initial recommender system. In this approach, a set of candidate recommendations that have been selected on the basis of relevance, are re-ranked in order to improve the diversity or novelty of the recommendation, or the aggregate diversity of all recommendations offered by the system. Generally, work that has taken this approach[4, 51, 65–67] attempts to optimize the set diversity as expressed by the ILD measure defined in Sect. 3.2.

In the recommendation context, a personalized recommendation is formed for a given target user u, and the relevance of any particular item to the recommendation depends on u. However, for notational simplicity, we will write $f_{rel}(i)$ for the relevance of item i, dropping the dependence on u. Given a candidate set C, the problem may be posed to find a set $R \subseteq C$ of some given size $k = |R|$, that maximizes $div(R)$ i.e.

$$R_{opt} = \underset{R \subseteq C, |R|=k}{\arg\max}\ div(R) \tag{8}$$

More generally, an objective to jointly optimize for relevance and diversity can be expressed as:

$$R_{opt}(\lambda) = \underset{R \subseteq C, |R|=k}{\arg\max}\ g(R, \lambda) \tag{9}$$

where

$$g(R, \lambda) = (1 - \lambda)\frac{1}{|R|}\sum_{i \in R} f_{rel}(i) + \lambda\, div(R)$$

and $\lambda \in [0, 1]$ expresses the trade-off between the average relevance of the items in the set and the diversity of the set. In Information Retrieval, a greedy construction approach to solving Eq. 9 is referred to as the maximum marginal relevance (MMR)

Algorithm 1 Greedy selection to produce a re-ranked list R from an initial set C

```
$R \leftarrow \emptyset$
while $|R| < k$ do
    $i* \leftarrow \arg\max_{i \in C-R} g(R \cup \{i\}, \lambda)$
    $R \leftarrow R \cup \{i*\}$
end while
return $R$
```

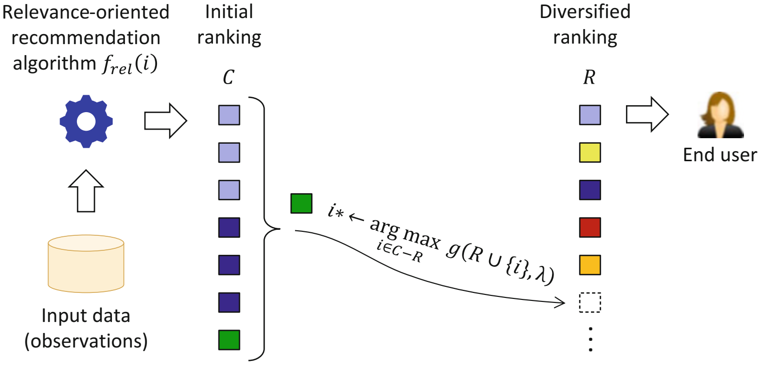

Relevance-oriented recommendation algorithm $f_{rel}(i)$
Input data (observations)
Initial ranking
C
$i* \leftarrow \arg\max_{i \in C-R} g(R \cup \{i\}, \lambda)$
Diversified ranking
R
End user

Fig. 2 Greedy reranking for a tradeoff between relevance and diversity

approach in [41], where relevance is measured with respect to a given query. In the greedy approach, the recommended set R is built in an iterative fashion as follows. Let R^j be the set at iteration $j \in \{1, 2, \ldots, k\}$. The first item in the set is the one that maximizes $f_{rel}(i)$ and the j-th item is chosen to maximize $g(R^{j-1} \cup \{i\}, \lambda)$. Algorithm 1 summarizes this approach, illustrated in Fig. 2.

In the context of case-based reasoning, a greedy solution to Eq. 9 is proposed in [2, 68] as a means of selecting a set of cases to solve a given target problem. Using nearest-neighbor user- and item-based collaborative filtering methods to generate the initial candidate set, Ziegler et al. [4] also propose a greedy solution to Eq. 9, as a means of re-ranking the set, terming the method as topic diversification, as they employ a taxonomy-based distance metric. In the context of a publish-subscribe system, Drosou and Pitoura [66] use the formulation in Eq. 9 as a means of selecting a diverse set of relevant items to recommend to a user from a set of items gathered over a particular time window. The method is proposed in the context of image retrieval in [65] and an alternative method to optimize for Eq. 9 is studied in [51], again using an item-based kNN method to generate the candidate set. Also, in [69] a number of different heuristics for solving the maximum diversity problem (Eq. 8) are evaluated and while none out-performs all others in all cases, several succeed in finding very good quality solutions in reasonable time. This work is followed up in [70], where a multiple-pass randomized greedy algorithm is shown to give better performance than the single-pass greedy algorithm.

Rather than maximize as a trade-off between relevance and diversity, [71] takes a more conservative approach of choosing the most diverse subset from a candidate set of items that have equal relevance, thereby maximizing diversity under a constraint of maintaining overall relevance. Similarly, [72] avoids using an explicit weighted trade-off between diversity and relevance and instead presents two algorithms that modify an initial relevance ranking of items to increase diversity.

Though it is difficult to compare directly across the different approaches, as the measures of relevance and pairwise distance differ, researchers have generally found the expected trade-off of increasing diversity and decreasing relevance of the retrieved set as λ is decreased towards 0. McGinty and Smyth [2, 68] evaluate the effect of diversifying the recommended set by counting the number of steps it takes a conversational recommender system to reach a given target item. Diversification always performs better than the algorithm that selects items using similarity only. An adaptive method that determines at each step whether or not to diversify gives even better performance. Evaluating on the Book-Crossing dataset, Ziegler et al. [4] found that the accuracy of their system, as measured by precision and recall, dropped with increasing diversification. Zhang and Hurley [51] evaluate on the Movielens dataset; they form test sets of increasing difficulty by splitting each user's profile into training and test sets of items, varying the average similarity of the items in the test set to the items in the training set, and find the diversified algorithm achieves better precision on the more difficult test sets.

Alternatives to MMR A number of alternative scoring functions for guiding re-ranking that capture the compromise between relevance and diversity or novelty have been proposed in the literature. For example, [73] computes a weighted sum of a global probability for including a candidate item in R and a local probability dependent on the set of items already in R. Definitions of novelty and diversity of news articles based on a distance between concepts in a taxonomy are given in [74] and a replacement heuristic is used to increase the novelty or diversity of the initial ranking by swapping in highly novel/diverse articles. To take account of mutual influence between items, [75] replace the pairwise diversity in the utility function by an estimate of the probability that the pair of items are both liked. Finally, an alternative formulation of the diversity problem encompassing both intra-list dissimilarity and external coverage, is presented in [76] and solved using a greedy algorithm. In this formulation, the items in the set R are selected to cover the candidate, such that there is an item in R within a certain similarity threshold to each item in C, under a constraint that all items in R also have a certain minimum pairwise dissimilarity.

Aggregate Diversity Targeting aggregate diversity (Eq. 3 in Sect. 3.5), items are re-ranked in [77] using a weighted combination of their relevance and a score based on inverse popularity, or item likeability. Adomavicius and Kwon find that their re-ranking strategies succeed in increasing aggregate diversity at a small cost to accuracy as measured by the precision. A follow-up to this work is presented in [6], in which the aggregate diversity problem is shown to be equivalent to the maximum flow problem on a graph whose nodes are formed from the users and items of the recommendation problem. More recently, Mansoury et al. [56] achieve significant aggregate diversity gains and a good accuracy tradeoff, by a maximum flow approach on the user-item bipartite graph defined by the rating matrix. Other work [78] has investigated how neighborhood filtering strategies and multi-criteria ratings impact on aggregate diversity in nearest-neighbor collaborative filtering algorithms. In this line, Vargas and Castells [58] find out that aggregate diversity is considerably improved by transposing the kNN CF recommendation approach, swapping the role of users and items. The authors show the approach can be generalized to any recommendation algorithm based on a probabilistic reformulation of arbitrary user-item scoring functions which isolates a popularity component.

4.2 Using Clustering for Diversification

A method proposed in [79] clusters the items in an active user's profile, in order to group similar items together. Then, rather than recommend a set of items that are similar to the entire user profile, each cluster is treated separately and a set of items most similar to the items in each cluster is retrieved.

A different approach is presented in [80], where the candidate set is again clustered. The goal now is to identify and recommend a set of representative items, one for each cluster, so that the average distance of each item to its representative is minimized.

A nearest-neighbor algorithm is proposed in [81] that uses multi-dimensional clustering to cluster items in an attribute space and select clusters of items as candidates to recommend to the active user. This method is shown to improve aggregate diversity.

A graph-based recommendation approach is described in [82] where the recommendation problem is formulated as a cost flow problem over a graph whose nodes are the users and items of the recommendation. Weights in the graph are computed by a biclustering of the user-item matrix using non-negative matrix factorization. This method can be tuned to increase the diversity of the resulting set, or increase the probability of recommending long-tail items.

4.3 Fusion-Based Methods

Since the early days of recommender systems, researchers have been aware that no single recommendation algorithm will work best in all scenarios. Hybrid systems have been studied to offset the strengths of one algorithm against the weaknesses of another (see [83] for example). It may be expected that the combined outputs of multiple recommendation algorithms that have different selection mechanisms, may also exhibit greater diversity than a single algorithm. For example, in [52, 84], recommendation is treated as a multi-objective optimization problem. The outputs of multiple recommendation algorithms that differ in their levels of accuracy, diversity and novelty are ensembled using evolutionary algorithms. As another example, in a music recommendation system called Auralist [33], a basic item-based recommender system is combined with two additional algorithms, in order to promote serendipity (see section below).

4.4 Incorporating Diversity in the Ranking Objective

Many recommender algorithms learn a model that optimises an objective representing a loss function that penalises mistakes in the ordering of the items by the model. Several works have examined ways to increase the diversity of the rankings produced by such models. For instance, [85] incorporates a diversity criterion in a regularisation term added to the loss function. In [86], the concept of diversity is integrated into a matrix factorization model, in order to directly recommend item sets that are both relevant and diversified. The loss function of the Bayesian Personalised Ranking (BPR) algorithm [87] is designed to distinguish a set of "relevant items" from a sampled set of "negative" items. Some works e.g. [88, 89] have examined ways to sample the negative items in order to enhance the diversity or novelty of the top k ranked items produced by the model. More recently, Li et al. [10] consider movie and music genre coverage (a diversity dimension) in the expected reward model—the objective function—for a multi-armed bandit recommendation algorithm.

4.5 Serendipity: Enabling Surprising Recommendations

A number of algorithms have been proposed in the literature to recommend serendipitous items. For example, in a content-based recommender system, described in [90], a binary classifier is used to distinguish between relevant and irrelevant content. Those items for which the difference in the positive and negative class scores is smallest are determined to be the ones about which the user is most uncertain and therefore the ones that are likely to yield serendipitous recommendations.

Oku and Hattori [91] propose a method for generating serendipitous recommendations that, given a pair of items, uses the pair to generate a recommended set of serendipitous items. Several ways to generate the set are discussed and several ways to rank the items and hence select a top k set are evaluated.

Utility theory is exploited in [5], where the utility of a recommendation is represented as a combination of its utility due to its quality and its utility due to its unexpectedness. A couple of different utility functions are proposed and ways to compute these functions on movie and book recommendation systems are discussed and evaluated.

Other work [92, 93] has investigated the use of graph-based techniques to make serendipitous recommendations in mobile app and music recommendation, respectively.

4.6 Other Approaches

A nearest neighbor algorithm called usage-context based collaborative filtering (UCBCF) is presented in [94], which differs from standard item-based CF in the calculation of item-item similarities. Rather than the standard item representation as a vector of user ratings, an item profile is represented as a vector of the k other items with which the item significantly co-occurs in user profiles. UCBCF is shown to obtain greater aggregate diversity than standard kNN and matrix factorization algorithms. A system described in [70] maps items into a utility space and maps a user's preferences to a preferred utility vector. In order to make a diverse recommendation, the utility space is split into m layers in increasing distance from the preferred utility and non-dominated items are chosen from each layer so as to maximize one dimension of the utility vector.

The works discussed so far have considered diversity in terms of the dissimilarity of items in a single recommendation set, or, in the case of aggregate diversity, the coverage of items in a batch of recommendations. Another approach is to consider diversity in the context of the behavior of the system over time. Temporal diversity (Eq. 6 in Sect. 3.5) is investigated by Lathia et al. [30] in a number of standard CF algorithms, and methods for increasing diversity through re-ranking or hybrid fusion are discussed. In a related vein, Mourão et al. [95] explore the "oblivion problem", that is, the possibility that in a dynamic system, items can be forgotten over time in such a way that they recover some degree of the original novelty value they had when they were discovered.

4.7 User Studies

It is one thing to develop algorithms to diversify top N lists, but what impact do these algorithms have on user satisfaction? A number of user studies have explored the impact of diversification on users. Topic diversification is evaluated in [4] by carrying out a user survey to assess user satisfaction with a diversified recommendation. In the case of their item-based algorithm, they find that satisfaction peaks around a relevance/diversity determined by $\lambda = 0.6$ in Eq. 9 suggesting that users like a certain degree of diversification in their lists.

While much of the work in diversifying top N lists does not consider the ordering of items in the recommended list, provided an overall relevance is attained, Ge et al.

[54, 96] look at how this ordering affects the user's perception of diversity. In a user study, they experiment with placing diverse items—ones with low similarity to the other items in the list—either in a block or dispersed throughout the list and found that blocking the items in the middle of the list reduces perceived diversity.

The work of Hu and Pu [97] addresses user-interface issues related to augmenting users' perception of diversity. In a user study that tracks eye movements, they find that an organizational interface where items are grouped into categories is better than a list interface in supporting the perception of diversity. In [98], 250 users are surveyed and presented with 5 recommendation approaches, with varying degrees of diversity. They find that users notice differences in diversity and diversity overall improves their satisfaction, but that diverse recommendations may require additional explanations to users who cannot link them back to their preferences.

More recently, in a larger experiment with over 2000 users (after data cleanup) on an e-commerce platform, Chen et al. [34] found serendipity to be a more decisive factor in user satisfaction than novelty or diversity alone. The latter were still important, combined with relevance, as factors of serendipity.

4.8 Diversification Approaches in Information Retrieval

Most diversification algorithms proposed in IR follow the same greedy re-ranking scheme as described earlier for recommender systems in Sect. 4.1. The algorithms distinguish from each other in the greedy objective function and the theory behind it. They can be classed into two types based on whether or not the algorithms use an explicit representation of query aspects (as introduced earlier in Sect. 3.8). Explicit approaches draw an approximation of query aspects from different sources, such as query reformulations suggested by a search engine [45], Wikipedia disambiguation entries [99], document classifications [40], or result clustering [100]. Based on this, the objective function of the greedy re-ranking algorithms seeks to maximize the number of covered aspects and minimize the repetition of aspects already covered in previous ranking positions. E.g. xQuAD [45], the most effective algorithm in TREC campaigns, defines its objective function as:

$$f(d_k|S,q) = (1-\lambda)\ p(q|d_k) + \lambda \sum_{z} p(z|q)\ p(d_k|q,z) \prod_{j=1}^{k-1} (1 - p(d_j|q,z))$$

where $p(q|d_k)$ stands for the initial search system score, z represents query aspects, $p(z|q)$ weights the contribution on each aspect by its relation to the query, $p(d_k|q,z)$ measures how well document d_k covers aspect z, the product after that penalizes the redundancy with previous documents in the ranking covering the same aspect, and $\lambda \in [0, 1]$ sets the balance in the intensity of diversification.

Diversification algorithms that do not explicitly deal with query aspects generally assess diversity in terms of the content of documents. For instance Goldstein and

Carbonell [41] greedily maximize a linear combination of similarity to the query (the baseline search score) and dissimilarity (minimum or average distance) to the documents ranked above the next document. Other non-aspect approaches formulate a similar principle in more formal probabilistic terms [43], or in terms of the trade-off between risk and relevance, in analogy to Modern Portfolio Theory [101] on the optimization of the expected return for a given amount of risk in financial investment. Vargas et al. [48, 49] show that IR diversity principles and techniques make sense in recommender systems and can be adapted to them as well, as we discuss in Sect. 5.4.

5 Unified View

As the overview through this chapter shows, a wide variety of metrics and perspectives have been developed around the same concepts under different variants and angles. It is natural to wonder whether it is possible to relate them together under a common ground or theory, establishing equivalences, and identifying fundamental differences. We summarize next a formal foundation for defining, explaining, relating and generalizing many different state of the art metrics, and defining new ones. We also examine the connections between diversity as researched and developed in the Information Retrieval field, and the corresponding work in recommender systems.

5.1 General Novelty/Diversity Metric Scheme

As shown in [9] it is possible indeed to formulate a formal scheme that unifies and explains most of the metrics proposed in the literature. The scheme posits a generic recommendation metric m as the expected novelty of the items it contains:

$$m = \frac{1}{|R|} \sum_{i \in R} nov(i|\theta)$$

An item novelty model $nov(i|\theta)$ at the core of the scheme determines the nature of the metric that will result. The scheme further emphasizes the relative nature of novelty by explicitly introducing a context θ. Novelty is relative to a context of experience: (what we know about) what someone has experienced somewhere sometime, where "someone" can be the target user, a set of users, all users, etc.; "sometime" can refer to a specific past time period, an ongoing session, "ever", etc.; "somewhere" can be the interaction history of a user, the current recommendation being browsed, past recommendations, recommendations by other systems, "anywhere", etc.; and "what we know about that" refers to the context of observation, i.e. the available observations to the system. We elaborate next on how such models can be defined, computed, and packed into different metrics.

5.2 Item Novelty Models

As discussed in Sect. 3.4, the novelty of an item can be established in terms of whether the item itself or its attributes have been experienced before. The first case, which we may refer to as an issue of simple item discovery, calls for a probabilistic formulation, whereas feature-based novelty, which we shall refer to as an issue of item familiarity, can be more easily defined in terms of a distance model.

5.2.1 Item Discovery

In the simple discovery approach, $nov(i|\theta)$ can be expressed in terms of the probability that someone has interacted with the item [9]. This probability can be defined from two slightly different perspectives: the probability that a random user has interacted with the item (to which we shall refer as forced discovery) as in IUF (Eq. 1), or the probability that the item is involved in a random interaction (free discovery). Both can be estimated based on the amount of interactions with the item observed in the system, as a sample of all the interactions the item may have received in the real world. We shall use the notation $p(known|i,\theta)$—the probability that "i is known" by any user given a context θ—for forced discovery, and $p(i|known,\theta)$ for free discovery. Note that these are different distributions, e.g. the latter sums to 1 over the set of all items, whereas the former sums to 1 with $p(\neg known|i,\theta)$. Forced discovery reflects the probability that a random user knows a specific item when asked about it, whereas free discovery is the probability that "the next item" someone discovers is precisely the given item. It is shown in [9] that the metrics induced by either model are quite equivalent in practice, as the two distributions are approximately proportional to each other (exactly proportional if the frequency of user-item pairs is uniform, as is the case e.g. with one-time ratings). In Sect. 7 we shall show some empirical results which confirm this near equivalence in practice.

Now depending on how we instantiate the context θ, we can model different novelty perspectives. For instance, if we take θ to be the set of available observations of user-item interaction (to be more rigorous, we take θ to be an unknown user-item interaction distribution of which the observed interactions are a sample), maximum-likelihood estimates of the above distributions yield:

$$\begin{aligned} p(known|i,\theta) &\sim |\{u \in \mathcal{U} \mid \exists t\ (u,i,t) \in \mathcal{O}\}| \ / \ |\mathcal{U}| = |\mathcal{U}_i|/|\mathcal{U}| \\ p(i|known,\theta) &\sim |\{(u,i,t) \in \mathcal{O}\}| \ / \ |\mathcal{O}| \end{aligned} \tag{10}$$

where $\mathcal{U}_i$ denotes the set of all users who have interacted with i, and $\mathcal{O} \subset \mathcal{U} \times \mathcal{I} \times \mathcal{T}$ is the set of observed item-user interactions with i (each labeled with a different timestamp $t \in \mathcal{T}$). If the observations consist of ratings, user-item pairs occur only once, and we have:

$$\begin{aligned} p(known|i,\theta) &\sim |\mathcal{U}_i|/|\mathcal{U}| = |\{u \in \mathcal{U} \mid r(u,i) \neq \emptyset\}| \;/\; |\mathcal{U}| \\ p(i|known,\theta) &\sim |\{u \in \mathcal{U} \mid r(u,i) \neq \emptyset\}| \;/\; |\mathcal{O}| = |\mathcal{U}_i|/|\mathcal{O}| \end{aligned} \tag{11}$$

Both $p(i|known,\theta)$ and $p(known|i,\theta)$ make sense as a measure of how popular an item is in the context at hand. In order to build a recommendation novelty metric based on this, we should take $nov(i|\theta)$ to be a monotonically decreasing function of these probabilities. The inverse probability, dampened by the logarithm function (i.e. $-\log_2 p$) is frequent in the literature [9, 28], but $1 - p$ is also reported as "popularity complement" [9, 52]. The latter has an intuitive interpretation when applied to forced discovery: it represents the probability that an item is not known to a random user. The former also has interesting connections: when applied to forced discovery, it gives the inverse user frequency IUF (see Sect. 3.3). When applied to free discovery, it becomes the self-information (also known as surprisal), an information theory measure that quantifies the amount of information conveyed by the observation of an event.

5.2.2 Item Familiarity

The novelty model scheme defined in the previous section considers how different an item is from past experience in terms of strict Boolean identity: an item is new if it is absent from past experience ($known = 0$) and not new otherwise ($known = 1$). There are reasons however to consider relaxed versions of the Boolean view: the knowledge available to the system about what users have seen is partial, and therefore an item might be familiar to a user even if no interaction between them has been observed in the system. Furthermore, even when a user sees an item for the first time, the resulting information gain—the effective novelty—ranges in practice over a gradual rather than binary scale (consider for instance the novelty involved in discovering the movie "Rocky V").

As an alternative to the popularity-based view, we consider a similarity-based model where item novelty is defined by a distance function between the item and a context of experience [9]. If the context can be represented as a set of items, for which we will intentionally reuse the symbol θ, we can formulate this as the distance between the item and the set, which can be defined as an aggregation of the distances to the items in the set, e.g. as the expected value:

$$nov(i|\theta) = \sum_{j \in \theta} p(j|\theta)\, d(i,j)$$

The $p(j|\theta)$ probability enables further model elaborations, or can be simply taken as uniform thus defining a plain distance average.

In the context of distance-based novelty, we find two useful instantiations of the θ reference set: (a) the set of items a user has interacted with—i.e. the items in his profile—, and (b) the set R of recommended items itself. In the first case, we get a user-relative novelty model, and in the second case, we get the basis for a generalization of intra-list diversity. The notion of expected set in [5] plays a similar role to this idea of θ context. It is possible to explore other possibilities for θ, such as groups of user profiles, browsed items over an interactive session, items recommended in the past or by alternative systems, etc., which might motivate future work.

5.3 Resulting Metrics

As stated at the beginning of this section, having defined a model of the novelty of an item, the novelty or diversity of a recommendation can be defined as the average novelty of the items it contains [9]. Each novelty model, and each context instantiation produce a different metric. In the following we show some practical instantiations that give rise to (hence unify and generalize) metrics described in the literature and covered in Sect. 3. Figure 3 informally illustrates the unified vision we develop in the next subsections.

5.3.1 Discovery-Based

A practical instantiation of the item discovery models described in Sect. 5.2.1 consists of taking the novelty context θ to be the set of user-item interactions

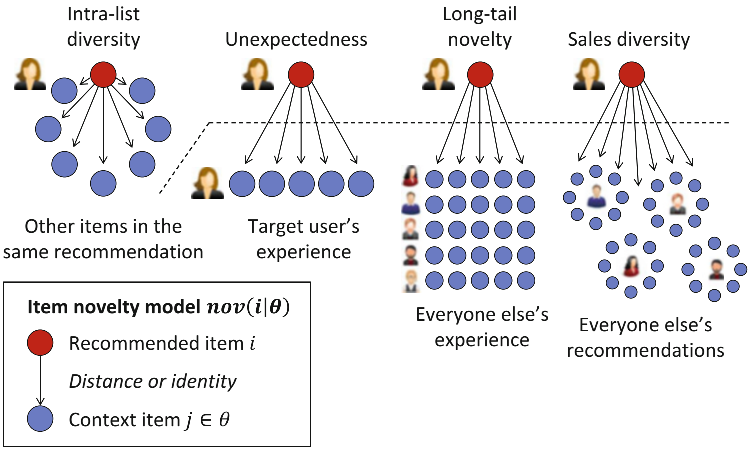

Intra-list diversity
Unexpectedness
Long-tail novelty
Sales diversity
Other items in the same recommendation
Target user's experience
Everyone else's experience
Everyone else's recommendations
Item novelty model $nov(i|\theta)$
Recommended item i
Distance or identity
Context item $j \in \theta$

Fig. 3 Different contexts of reference for item novelty reflect different diversity and novelty notions and result in different metrics

observed by the system. The different discussed variants in the novelty model result in the following practical metric combinations (mean IUF, mean self-information, mean popularity complement):

$$\begin{aligned}
\text{MIUF} &= -\frac{1}{|R|}\sum_{i\in R}\log_2 p(known|i,\theta)\\
\text{MSI} &= -\frac{1}{|R|}\sum_{i\in R}\log_2 p(i|known,\theta) \qquad (12)\\
\text{MPC} &= \frac{1}{|R|}\sum_{i\in R}(1-p(known|i,\theta))
\end{aligned}$$

where the probabilities are estimated by Eqs. 10 or 11 depending on the nature of the data. MPC has the advantage of simplicity, a clear interpretation (the ratio of unknown recommended items), and ranges in [0, 1]. MIUF generalizes the metric proposed by Zhou et al. [28] (Eq. 1 in Sect. 3.3), and MSI provides a nice connection to information theory concepts. MPC has the potential shortcoming of a tendency to concentrate its values in a small range near 1, whereas MIUF and MSI deliver less clumped values. We might as well consider the expected popularity complement of free discovery, but that does not have a particularly interesting interpretation or property with respect to the other metrics. In fact, given the discussed near equivalence of free and forced discovery, the three above metrics behave quite similarly to each other, as we will illustrate in Sect. 7 for MSI and MPC.

5.3.2 Familiarity-Based

Distance based item novelty models give rise to intra-list diversity and unexpectedness metrics. As mentioned in Sect. 5.2.2, these metrics simply result from taking, respectively, the recommended items or the target user's profile as the θ novelty context. The complement of any similarity function between items is potentially suitable to define the distance measure. For instance, with feature-based similarity we may define $d(i, j) = 1 - \cos(i, j)$ for numeric item features, $d(i, j) = 1 - \text{Jaccard}(i, j)$ for Boolean (or binarized) features, and so forth. The distinction between collaborative and content-based similarity deserves attention though, and care should be taken to make a meaningful choice between these two alternatives. Content-based similarity compares items by their intrinsic properties, as described by the available item features. Even though a collaborative similarity measure (which compares items by their common user interaction patterns) might make sense in some particular cases, we would contend that content-based similarity is generally more meaningful to assess the diversity in a way that users can perceive.

5.3.3 Further Unification

By explicitly modeling novelty as a relative notion, the proposed framework has a strong unifying potential of further novelty and diversity conceptions. Take for instance the notion of temporal diversity [30] discussed in Sect. 3.5. The metric can be described in the framework in terms of a discovery model where the source of discovery is the past recommendations of the system $\theta \equiv R'$, and novelty is defined as the complement of forced discovery given this context:

$$\frac{1}{|R|}\sum i \in R(1 - p(known|i, R') = \frac{1}{|R|}\sum_{i \in R}(1 - [i \in R']) = \frac{1}{|R|}|R - R'| = \text{TD}$$

Similarly, for inter-user diversity (Eq. 4 in Sect. 3.5), we take as context the set of recommendations to all users in the system, $\theta \equiv \{R_v | v \in \mathcal{U}\}$. By marginalization over users, and assuming a uniform user prior $p(v) = 1/|\mathcal{U}|$ we have:

$$\frac{1}{|R|}\sum_{i \in R}(1 - p(known \mid i, \{R_v | v \in \mathcal{U}\}) = \frac{1}{|R|}\sum_{i \in R}\sum_{v \in \mathcal{U}}(1 - p(known|i, R_v))p(v)$$

$$= \frac{1}{|R||\mathcal{U}|}\sum_{v \in \mathcal{U}}\sum_{i \in R}(1 - p(known|i, R_v)) = \frac{1}{|R||\mathcal{U}|}\sum_{v \in \mathcal{U}}|R - R_v| = \text{IUD}$$

Inter-system novelty can be obtained in a similar way. So can generalized unexpectedness metrics, in their set difference form (Eq. 2 in Sect. 3.4), by using the expected set as context θ in place of R', R_v or R_s above.

Biodiversity measures from ecology can also be directly related to some of the recommendation metrics we have discussed. The equivalences hold by equating items to species, and the occurrence of an item in a recommendation as the existence of an individual of the species. In particular, stated in this way, aggregate diversity is the direct equivalent of so called richness, the number of different species that are present in an ecosystem [38, 39].

On the other hand, it can be seen that the Gini-Simpson index (GSI) is exactly equivalent to inter-user diversity. GSI is defined as the probability that two items (individuals) picked at random from the set of recommendations (ecosystem) are different items (species) [38, 39], which can be expressed as a sum over items, or as an average over pairs of recommendations:

$$\text{GSI} = 1 - \sum_{i \in \mathcal{I}}\frac{|\{u \in \mathcal{U} \mid i \in R_u\}|^2}{|\mathcal{U}|(|\mathcal{U}| - 1)k^2} = 1 - \frac{1}{|\mathcal{U}|(|\mathcal{U}| - 1)}\sum_{u \in \mathcal{U}}\sum_{v \in \mathcal{U}}\frac{|R_u \cap R_v|}{|R_u||R_v|}$$

where $k = |R_u|$ assuming they are the same size, or equivalently, considering we are computing GSI@k, and we assume item pairs are not sampled from the same recommendation. On the other hand, the average value of IUD over all users is:

Table 1 Some item novelty and context model instantiations, and the metric they result into

Metric	Used in	Item novelty model	Context
ILD	[2, 4, 9, 51, 52]	$\sum_{j \in \theta} p(j \mid \theta)\, d(i, j)$	$\theta \equiv R$
Unexp	[5, 9, 32, 54]		$\theta \equiv$ items in user profile
MIUF	[28]	$-\log_2 p(known \mid i, \theta)$	$\theta \equiv$ all observed user-item interaction data
MSI	[9]	$-\log_2 p(i \mid known, \theta)$	
MPC	[9, 52, 84]		
TD	[30]	$1 - p(known \mid i, \theta)$	$\theta \equiv$ items recommended in the past
IUD/GSI	[28]		$\theta \equiv \{R_u \mid u \in \mathcal{U}\}$
ISD	[29]		$\theta \equiv \{R^s \mid s \in \mathcal{S}\}$

$$\text{IUD} = \frac{1}{|\mathcal{U}|(|\mathcal{U}| - 1)} \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{U}} \frac{|R_u - R_v|}{|R_u|} = 1 - \frac{1}{|\mathcal{U}|(|\mathcal{U}| - 1)} \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{U}} \frac{|R_u \cap R_v|}{|R_u|}$$
$$= 1 - k(1 - \text{GSI}) \propto \text{GSI}$$

□

Table 1 summarizes some of the metrics that can be obtained in the unified framework by different instantiations of θ and item novelty models.

5.3.4 Direct Optimization of Novelty Models

The metric scheme described in the previous sections enables the definition of novelty or diversity enhancement re-ranking methods by the greedy optimization of an objective function combining the initial ranking score and the novelty model value, as described earlier in Sect. 4.1:

$$g(i, \lambda) = (1 - \lambda) f_{rel}(i) + \lambda\, nov(i \mid \theta)$$

By taking a particular novelty model $nov(i \mid \theta)$, one optimizes for the corresponding metric that takes the model at its core. This is an approach to diversity enhancement which is by definition difficult to overcome—in terms of the target metrics—by other re-ranking means.

5.4 Connecting Recommendation Diversity and Search Diversity

Recommendation can be formulated as an Information Retrieval task, one where there is no explicit user query. To this extent, and in the aim to find a perspective as comprehensive as possible on the topic at hand in this chapter, it is natural to wonder whether it is possible to establish a connection between the work on diversity in both fields. This question finds affirmative answers in many senses [9, 48, 49]. We summarize here what we find to be the main considerations in this direction: (a) recommendation novelty and diversity can be extended to be sensitive to relevance and rank, (b) IR diversity principles, metrics and algorithms can be adapted to a recommendation setting, and c) personalized search diversity can be formalized as a link between search and recommendation diversity.

5.4.1 Rank and Relevance

The novelty and diversity metrics described so far generally lack two aspects: they consider neither the relevance nor the rank position of the items when assessing their contribution to the novelty value of the recommendation. This is in contrast to IR metrics such as ERR-IA and α-nDCG which add up the novelty contribution of items only when they are relevant to the query, and apply a rank discount reflecting the assumption that lower ranking positions are less likely to be actually reached by users. Some authors in the recommender systems domain have also proposed to take relevance into account [8, 28, 32, 54, 60], though it is most often not the case, and rank position is generally not taken into account in the reported metrics.

Vargas and Castells [9] show that it is possible to deal with relevance and novelty or diversity together by introducing relevance as an intrinsic feature to the unified metric scheme described in the previous section. This can be done by just replacing “average” by “expected” item novelty at the top level of the scheme, where the novelty of a recommended item should only count when it is actually seen and consumed (chosen, accepted) by the user. The expected novelty is then computed in terms of the probability of choice. If we make the simplifying assumptions that (a) the user chooses an item if and only if she discovers it and likes it, and (b) discovery and relevance are independent, the resulting scheme is:

$$m = C \sum_{i \in R} p(seen|i, R)\ p(rel|i)\ nov(i|\theta) \tag{13}$$

where $p(rel|i)$ estimates the probability that i is relevant for the user, achieving the desired effect that only relevant novel items count, and $p(seen|i, R)$ estimates the probability that the user will get to see the item i while browsing R.

The probability of relevance can be defined based on relevance judgments (test ratings), for instance as $p(rel|i) \sim r(u, i)/r_{max}$, where r_{max} is the maximum possible rating value. Assessing relevance and diversity together has several advantages. It allows for a unified criteria to compare two systems, where separate relevance and novelty metrics may disagree. Furthermore, assessing relevance and novelty together allows distinguishing, for example, between recommendations A and B in Table 2: B can be considered better (relevance-aware MPC = 0.5) since it recommends one useful item (relevant and novel), whereas the items recommended by A (relevance-aware MPC = 0) lack either relevance or novelty. Note that

Table 2 Toy example recommendations of size two by three systems A, B, C. For each item, the pairs of check and cross marks indicate whether or not the item is relevant (left) and novel (right) to the user (e.g. item 1 of A is relevant and not novel). Below this, the values of MPC are shown with different combinations of rank discount and relevance awareness: plain MPC without relevance or rank discounts, relevance-weighted MPC (without rank discount), and MPC with a Zipfian rank discount (without relevance). The specific expression of the discount function $p(seen|i_k, R)$ and the relevance weight $p(rel|i)$ is shown for each metric variant. The last two rows show the precision of each recommendation, and the harmonic mean of precision and plain MPC

Rank	**A**	**B**	**C**
1	✓✗	✓✓	✗✗
2	✗✓	✗✗	✓✓

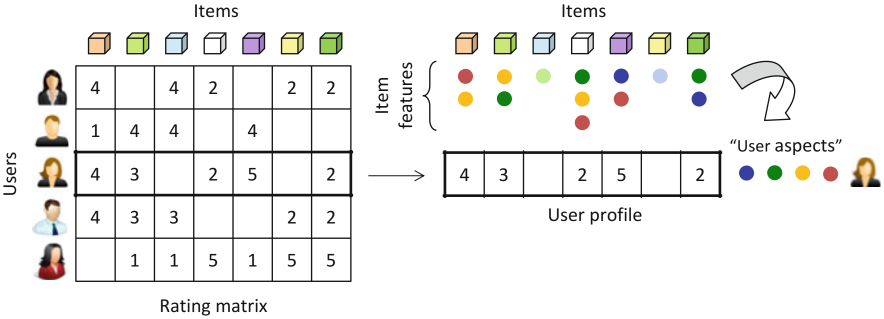
Metric	$p(seen \mid i_k, R)$	$p(rel \mid i)$	**A**	**B**	**C**
Plain MPC	1	1	1	0.5	0.5
Relevance-aware MPC	1	$r(u,i)/r_{max}$	0	0.5	0.5
Zipfian MPC	$1/k$	1	0.25	0.5	0.25
Precision	1	$r(u,i)/r_{max}$	1	0.5	0.5
H(Plain MPC, Precision)	–	–	1	0.5	0.5

an aggregation of separate novelty and relevance metrics would not catch this difference—e.g. the harmonic mean of MPC and precision is 0.5 for both A and B.

On the other hand, the $p(seen|i, R)$ distribution allows the introduction of a browsing model of a user interacting with the ranked recommendations, thus connecting to work on the formalization of utility metrics in IR [42, 102, 103]. The browsing model results in a rank discount which reflects the decreasing probability that the user sees an item as she goes down the ranking. Different models result in discount functions such as logarithmic $p(seen|i_k, R) = 1/\log_2 k$ as in nDCG (see Chap. 15), exponential p^{k-1} as in RBP [103], Zipfian $1/k$ as in ERR [42], and so forth (see [102] for a good compendium and formalization of alternatives). Rank discount allows distinguishing between recommendations B and C in Table 2: B is better (Zipfian MPC $= 0.5$) since it ranks the relevant novel item higher than C (Zipfian MPC $= 0.25$), with higher probability to be seen by the user.

5.4.2 IR Diversity in Recommendation

Vargas et al. [48, 49] have shown that the IR diversity principles, metrics and algorithms can be directly applied to recommendation. At a theoretical level, the evidence of user needs implicit in user actions is generally even more ambiguous and incomplete to a recommender system than an explicit query can be to a search system, whereby the rationale of diversifying retrieved results to increase the chances of some relevant result applies here as well.



4		4	2		2	2
1	4	4		4		
4	3		2	5		2
4	3	3			2	2
	1	1	5	1	5	5

4	3		2	5		2

Fig. 4 User aspects for diversification, as an analogous to query aspects, can be derived from item features and user ratings

At a practical level, it makes as much sense to consider the different aspects of a user's preferences (there are different sides to a person's interests) as it can make for an expressed query. A user interest aspect representation can be drawn from item features in some suitable, meaningful space. This can be done in analogy to the document categories as handled in [40]. Figure 4 illustrates the idea. From this point on, IR diversity metrics such as ERR-IA [42] or subtopic recall [46] can be applied, and aspect-based algorithms such as xQuAD [45] or IA-Select [40] can be adapted, equating users to queries and items to documents. In [104, 105], the aspect representation is explored further, by expressing the trade-off between accuracy and diversity in objectives other than those imported directly from IR, such as the variance minimization objective of modern portfolio theory, as well as proposing an algorithm to learn the aspect probabilities on which these objectives depend.

Non-aspect based diversification methods are applied even more straightforwardly to recommendation, as proved by the equivalence between MMR [41] and methods in the recommender systems field [2, 4] (Sect. 4.1), or the adaptation of the Modern Portfolio Theory from IR [101] to recommendation [82].

5.4.3 Personalized Diversity

Recommender and search systems can be seen as extremes in the explicit vs. implicit spectrum of the available evidence of user needs: a recommender system takes no explicit query and relies on observed user choices—implicit evidence of user preferences—as input, whereas basic search systems use just an explicit query. Personalized search represents a middle ground in this spectrum, using both an explicit user query and implicit user feedback and observed actions.

The possibility to consider diversity in the presence of both a query and a user profile has been researched as well [106, 107]. This standpoint has interesting philosophical implications, since personalization can also be seen as a strategy to cope with the uncertainty in a user query. While in a diversification approach the system accepts a situation of uncertainty and adapts its behavior to it, personalization tries to reduce the uncertainty by enhancing the system knowledge about the user need.

Diversity and personalization do not necessarily exclude each other, and can in fact be combined into personalized diversification, as shown in [107]. Vallet and Castells developed and tested a full framework that generalizes IR diversification methods (including xQuAD [45] and IA-Select [40]) into a personalized diversification scheme, by introducing the user as a random variable in the probabilistic formalization of the diversity algorithm [107]. In addition to bridging two theories, the scheme compared favorably empirically to personalization and diversity alone.

6 Bias and Fairness

As counterparts of diversity from different perspectives, bias [108–111] and unfairness [112–116] have raised growing concern over the last decade in the recommender systems field, as well as wider disciplines such as Artificial Intelligence or Information Retrieval [117–120]. Fairness in recommendation is covered in Chap. 18, and the reader is referred there for a more comprehensive survey. Bias in recommendation has further angles besides fairness and an in-depth review is beyond the scope of the present chapter We briefly discuss nonetheless here some ways in which novelty, diversity, bias and fairness relate to each other.

6.1 Bias in Recommendation

In the context of recommendation, bias can refer to algorithms or to evaluation, and commonly works in opposition to novelty and diversity. A system is said to be biased when it is *systematically* inclined to recommend some items over others, for causes that are—a priori—unrelated to the purpose of recommendation: bringing value to the involved stakeholders (customers, providers, etc.). Seen as a matter of uneven distributions over items this is usually referred to as the *popularity bias*: different researchers have found [109, 121] and explained why [122] many collaborative filtering approaches are structurally biased to concentrate recommendations on majority taste (best sellers, popular trends, etc.), to the detriment of long-tail novelty and sales diversity. The recommendation feedback loop, if not properly handled, adds to this concentration effect [35, 123].

Bias in evaluation means that an experimental procedure is *systematically* favoring some algorithmic approaches over others. Evaluation bias is particularly strong in offline evaluation and is essentially introduced in the sampling of test data for metric evaluation. If test data is sampled from a non-uniform distribution over user-item pairs, algorithms are rewarded for not only predicting user tastes, but also for guessing where the test data is placed—which is irrelevant when assessing the system effectiveness. Ranking-based offline evaluation [108, 110, 111] is commonly exposed to strong item popularity biases, rewarding the behavior of popularity-biased algorithms and encouraging their deployment. From the realization of such biases, new algorithmic approaches [124–126], metrics, and experimental procedures [108, 127–130] have been researched to better cope with biases.

Dealing with popularity biases and caring for novelty are related but different perspectives: the former aims to avoid distortion in evaluating for relevance, while the latter cares about user satisfaction as a direct result of novel experiences, beyond the most popular ones. Interestingly, unexplored equivalences lay between the two perspectives. To see this, let us consider *Inverse propensity scoring* (IPS) [131], a debiasing technique, that is receiving increasing attention in this area [128–130]. IPS corrects for the bias in evaluation by dividing relevance by the "propensity" of relevance to be observed in the computation of offline metrics [129, 130]. For instance, the corrected version of a metric such as precision is:

$$\mathrm{P_{ips}} = \frac{1}{|R|} \sum_{i \in R} \frac{rel_i}{p(obs|i)} \tag{14}$$

where $rel_i = 1$ if i is relevant and zero otherwise, and we make the target user implicit (i.e. the above definition computes precision of a single recommendation, to be averaged then over all users). The term $p(obs|i)$ represents propensity: the probability that the target user is observed interacting with item i as part of the test data for evaluation.

Taking popularity as a simplest, user-independent propensity estimate (as in e.g. [108, 130]) IPS is almost equivalent to relevance-aware MIUF. Combining Eqs. 12 and 13, taking a flat browsing model $p(seen|i, R) = 1$, binary relevance $p(rel|i) = rel_i$, and a typical normalizing constant $C = 1/|R|$, we have:

$$\begin{aligned} \mathrm{MIUF} &= -\frac{1}{|R|} \sum_{i \in R} rel_i \log_2 p(known|i, \theta) \\ &= \frac{1}{|R|} \sum_{i \in R} rel_i \log_2 \frac{1}{p(known|i, \theta)} \end{aligned} \tag{15}$$

If we take a common popularity-based discovery model such as Eqs. 10 or 11 in both the item novelty model and item propensity: $p(known|i, \theta) \equiv |\mathcal{U}_i|/|\mathcal{U}| \equiv p(obs|i)$, we can see the close analogy between Eqs. 14 and 15 above. They are the same except for the logarithm (a monotonic function) in MIUF. In fact, this logarithm could be seen as a form of dampening growth to cope with the well-known high variance problem of IPS and the overdominance of items with smallest $p(obs|i)$, for which modified versions of IPS are often in fact used [127, 128].

We can find sense in this connection between novelty and bias compensation: novelty aims to reward what the user is less likely familiar with, whereas techniques such as IPS aim to reward the recommendation of choices that are less likely to be observed in evaluation. Of course, both conditions are related, since observations of unusual user experiences naturally tend to be scarce in evaluation data.

6.2 Fair Recommendation

Biases and poor diversity may have further consequences than a statistical distortion in evaluation or a stale end-user experience. Recommending some items much more frequently than others (low "sales diversity") can be unfair when different providers and creators are involved behind different items (artists on Spotify, vendors on Amazon, owners and operators in the travel and leisure industry, candidates in job recommendation, people in social networks, book authors, research authors, etc.) [112, 114, 115, 132]. Failing to reflect different options, viewpoints or plural opinions (e.g. in recommended news or readings) may work against freedom of choice, critical thinking and healthy societies [24]. In these perspectives, diversity and novelty may not just be a means for business optimization but also an issue of fair opportunity and ethical concern [115, 117, 120, 133].

Fairness can be viewed as a particular case of diversity and as such, some of the views described in the previous sections can be adapted or elaborated upon to measure and enhance different forms of fairness. For instance, fair opportunity can be represented as a particularization and/or elaboration of sales diversity principles discussed in Sect. 3.5. Item vendors, vendor data, news polarity, and other sensitive item features can be represented as item aspects as discussed in Sect. 5.4.2, and handled as diversity objectives [113, 115, 116]. Chapter 18 reviews these and related perspectives in more depth.

7 Empirical Metric Comparison

We illustrate the metrics and some of the algorithms described along this chapter with some empirical measurements for a few recommendation algorithms on MovieLens 1M. In the tests we present, ERR-IA and subtopic recall are defined using movie genres as user aspects; ILD and unexpectedness take Jaccard on genres as the distance measure; and aggregate diversity is presented as a ratio over the total number of items, for a better appreciation of differences.

Table 3 shows the metric values for some representative recommendation algorithms: matrix factorization (MF) recommender, based on [134]; a user-based kNN algorithm using the Jaccard similarity, and omitting the normalization by the sum of similarities in the item prediction function; a content-based recommender using movie tags; a most-popular ranking; and random recommendation. We may mainly notice that matrix factorization stands out as the most effective in aggregate diversity and ERR-IA. The latter can be attributed to the fact that this metric takes

Table 3 Novelty and diversity metrics (at cutoff 10) on a few representative recommendation algorithms in the MovieLens 1M dataset. The highest value of each metric is shown in boldface and table cells are colored in shades of green, darker representing higher values (the values of random recommendation are disregarded in the color and bold font of the rest of rows—though not vice versa—to allow the appreciation of differences excluding the random recommendation)

	nDCG	ILD	Unexp	MSI	MPC	Aggdiv	IUD	Entropy	ERR-IA	S-recall
MF	**0.3161**	0.6628	0.7521	9.5908	0.8038	**0.2817**	**0.9584**	**8.5906**	**0.2033**	0.5288
u-kNN	0.2856	0.6734	0.7785	9.0716	0.7361	0.1589	0.8803	7.1298	0.1800	**0.5422**
CB	0.1371	**0.6825**	0.7880	**9.7269**	**0.8101**	0.1650	0.7762	6.2941	0.1001	0.5378
PopRec	0.1415	0.6624	**0.8451**	8.5793	0.6514	0.0183	0.4943	4.5834	0.0773	0.5253
Random	0.0043	**0.7372**	0.8304	**13.1067**	**0.9648**	**0.9647**	**0.9971**	**11.7197**	0.0034	0.5055

relevance into account, a criteria on which MF achieves the best results of this set (as seen in nDCG).

Content-based recommendation procures the best long tail novelty metrics, confirming a well-known fact [7]. It is not comparably as bad in unexpectedness as one might expect, and this can be attributed to the fact that movie genres (the basis for the unexpectedness distance) and movie tags (the basis for CB similarity) seem not to correlate that much. This, and the good results in terms of ILD can also be related to the suboptimal accuracy of this algorithm as a standalone recommender, which may lend it, albeit to a small degree, some of the flavor of random recommendations. We have checked (outside the reported results) that CB naturally gets the lowest ILD value of all recommenders if the metric uses the same features as the CB algorithm (i.e. movie tags).

Popularity has (almost by definition) the worst results in terms of most novelty and diversity metrics; except in terms of unexpectedness (distance to user profile), which makes sense since this algorithm ignores any data of target users and thus delivers items that are weakly related to the individual profiles. Random recommendation is naturally optimal at most diversity metrics except the one that takes relevance into account (it has low subtopic recall though, because of a bias in MovieLens whereby genre cardinality—therefore subtopic coverage—correlates negatively with popularity). And kNN seems to achieve a good balance of the different metrics. We may also notice that aggregate diversity, IUD and entropy go hand in hand, as one would expect.

In order to give an idea of how related or different the metrics are, Table 4 shows the pairwise Pearson correlation of the metrics on a user basis for the MF recommender. We see that ILD, unexpectedness and subtopic recall tend to go hand in hand, even though they capture different properties as seen previously in the comparison of recommenders in Table 3 (e.g. popularity has very good unexpectedness but very poor ILD). MSI and MPC confirm to be quite equivalent, and IUD (which is equivalent to Gini-Simpson) goes strongly along with these long tail metrics. Note that aggregate diversity and entropy do not have a definition for individual users, and therefore they cannot be included in this table. However, as mentioned before, these measures show strong system-wise correspondence with IUD in Table 3 and, by

Table 4 Pearson correlation between different metrics (on a user basis) applied to a matrix factorization algorithm on MovieLens 1M. The shades of color (red for negative, green for positive) highlight the magnitude of values

nDCG							
0.64	**ERR-IA**						
0.03	-0.02	**S-recall**					
0.03	-0.09	0.71	**ILD**				
0.07	-0.06	0.62	0.85	**Unexp**			
0.02	0.09	-0.21	-0.21	-0.20	**MSI**		
0.02	0.10	-0.19	-0.21	-0.19	0.97	**MPC**	
0.06	0.14	-0.20	-0.27	-0.23	0.87	0.93	**IUD**

Table 5 Novelty and diversity metrics (at cutoff 10) on a few novelty and diversity enhancement algorithms applied to the matrix factorization algorithm on MovieLens 1M. The diversifiers are denoted either by their common name, or by the name of the metric (the item novelty model) they target in their objective function

	nDCG	ILD	Unexp	MSI	MPC	Aggdiv	ERR-IA	S-recall
MF	**0.3161**	0.6628	0.7521	9.5908	0.8038	0.2817	0.2033	0.5288
+MMR	0.2817	**0.7900**	0.8089	9.6138	0.8054	0.2744	0.1897	**0.6814**
+Unexp	0.2505	0.7588	**0.8467**	9.6011	0.8029	0.2483	0.1431	0.6439
+MSI	0.2309	0.6130	0.7384	**10.6995**	**0.8961**	**0.4700**	0.1583	0.4483
+MPC	0.2403	0.6233	0.7389	10.3406	0.8818	0.3683	0.1622	0.4696
+xQuAD	0.2726	0.6647	0.7596	9.5784	0.8034	0.2292	**0.2063**	0.6370
+Random	0.0870	0.6987	0.7698	10.2517	0.8670	**0.4836**	0.0623	0.5561

transitivity, can be expected to correlate with long-tail metrics as well. The correlation between ERR-IA and nDCG reflects the fact that in addition to aspect diversity ERR-IA takes much into account relevance, which is what nDCG measures.

Finally, and just for the sake of illustration, we see in Table 5 the effect of different novelty/diversity enhancers, applied to the best performing baseline in nDCG, namely matrix factorization. The diversifiers labeled as MMR, Unexp, MSI and MPC are greedy optimizations of the corresponding item novelty model of each metric. xQuAD is an implementation of the algorithm described in [45] using movie genres as aspects, an algorithm which implicitly targets ERR-IA. We arbitrarily set $\lambda = 0.5$ for all algorithms, without a particular motive other than illustrative purposes. We can see that each algorithm maximizes the metric one would expect. The fact that MSI appears to optimize MPC better than MPC itself is because (a) both metrics are almost equivalent, and (b) $\lambda = 0.5$ is not the optimal value for optimization, whereby a small difference seems to tip the scale towards MSI by pure chance. Please note that these results are in no way aiming to approximate optimality or evaluate an approach over another, but rather to exemplify how the different models, metrics and algorithms may work and relate to each other in a simple experiment.

8 Conclusion

The consensus is clear in the community on the importance of novelty and diversity as fundamental qualities of recommendations. They are seen today as natural components in the continued improvement and evolution of recommendation technology. Considerable progress has been achieved in the area in defining novelty and diversity from several points of view, devising methodologies and metrics to evaluate them, and developing different methods to enhance them. This chapter aims to provide a wide overview on the work so far, as well as a unifying perspective linking them together as developments from a few basic common root principles.

The area still has wide space for further research. There is room for improving our understanding of the role of novelty and diversity in recommendation, and innovating in theoretical, methodological and algorithmic developments around these dimensions. For instance, modeling feature-based novelty in probabilistic terms in order to unify discovery and familiarity models would be an interesting line for future work. Aspects such as the time dimension, along which items may recover part of their novelty value [95, 135, 136], or the variability among users regarding their degree of novelty-seeking trend [14], are example directions for additional research. Last but not least, continued user studies [4, 15–17, 137] would bring further light on questions such as whether novelty and diversity metrics match the actual user perceptions, the precise extent and conditions in which users appreciate novelty and diversity versus accuracy and other potential dimensions of recommendation effectiveness.

Acknowledgments Neil Hurley would like to acknowledge the support of Science Foundation Ireland, under grant no. SFI/12/RC/2289_P2. Pablo Castells was supported by the Spanish Ministry of Science and Innovation, grant no. PID2019-108965GB-I00.

References

1. J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)
2. B. Smyth, P. McClave, in *Proceedings of the 4th International Conference on Case-Based Reasoning*, ICCBR 2001 (Springer, London, UK, 2001), pp. 347–361
3. S.M. McNee, J. Riedl, J.A. Konstan, in *CHI 2006 Extended Abstracts on Human Factors in Computing Systems*, CHI EA 2006 (ACM, New York, NY, USA, 2006), pp. 1097–1101
4. C.N. Ziegler, S.M. McNee, J.A. Konstan, G. Lausen, in *Proceedings of the 14th International Conference on World Wide Web*, WWW 2005 (ACM, New York, NY, USA, 2005), pp. 22–32
5. P. Adamopoulos, A. Tuzhilin, Special Issue on Novelty and Diversity in Recommender Systems, ACM Trans. Intell. Syst. Tech. **5**(4) (2014)
6. G. Adomavicius, Y. Kwon, IEEE Trans. Knowl. Data Eng. **24**(5), 896–911 (2012)
7. O. Celma, P. Herrera, in *Proceedings of the 2nd ACM Conference on Recommender Systems*, RecSys 2008 (ACM, New York, NY, USA, 2008), pp. 179–186
8. N. Hurley, M. Zhang, ACM Trans. Internet Tech. **10**(4), 14:1–14:30 (2011)

9. S. Vargas, P. Castells, in *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys 2011 (ACM, New York, NY, USA, 2011), pp. 109–116
10. C. Li, H. Feng, M.d. Rijke, in *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys 2020 (ACM, New York, NY, USA, 2020), pp. 33–42
11. P. Li, A. Tuzhilin, ACM Trans. Intell. Syst. Tech. **11**(6) (2020)
12. P. Li, M. Que, Z. Jiang, Y. HU, A. Tuzhilin, in *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys 2020 (ACM, New York, NY, USA, 2020), pp. 279–288
13. M. Kaminskas, D. Bridge, ACM Trans. Inter. Intell. Syst. **7**(1) (2016)
14. K. Kapoor, V. Kumar, L. Terveen, J.A. Konstan, P. Schrater, in *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys 2015 (ACM, New York, NY, USA, 2015), pp. 19–26
15. B.P. Knijnenburg, M.C. Willemsen, Z. Gantner, H. Soncu, C. Newell, User Model. User Adap. Inter. **22**(4-5), 441–504 (2012)
16. P. Pu, L. Chen, R. Hu, in *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys 2011 (ACM, New York, NY, USA, 2011), pp. 157–164
17. M.C. Willemsen, M.P. Graus, B.P. Knijnenburg, User Model. User Adap. Inter. **26**(4), 347–389 (2016)
18. L. McAlister, E.A. Pessemier, J. Consum. Res. **9**(3), 311–322 (1982)
19. S.R. Maddi, in *Theories of Cognitive Consistency: A Sourcebook*, ed. by R.P. Abelson, E. Aronson, W.J. McGuire, T.M. Newcomb, M.J. Rosenberg, P.H. Tannenbaum (Rand McNally, 1968)
20. P.S. Raju, J. Consum. Res. **7**(3), 272–282 (1980)
21. C. Coombs, G.S. Avrunin, Psychological Review **84**(2), 216–230 (1977)
22. P. Brickman, B. D'Amato, J. Pers. Soc. Psychol. **32**(3), 415–420 (1975)
23. B.E. Kahn, J. Intell. Inf. Syst. **2**(3), 139–148 (1995)
24. E. Pariser, *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* (Penguin Books, 2012)
25. M. Lubatkin, S. Chatterjee, Acad. Manag. J. **37**(1), 109–136 (1994)
26. C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More* (Hyperion, 2006)
27. G. Adomavicius, Y. Kwon, INFORMS J. Comput. **26**(2), 351–369 (2014)
28. T. Zhou, Z. Kuscsik, J.G. Liu, M. Medo, J.R. Wakeling, Y.C. Zhang, Proc. Natl. Acad. Sci. **107**(10), 4511–4515 (2010)
29. A. Bellogín, I. Cantador, P. Castells, Information Sciences **221**, 142–169 (2013)
30. N. Lathia, S. Hailes, L. Capra, X. Amatriain, in *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2010 (ACM, New York, NY, USA, 2010), pp. 210–217
31. Y.J. Park, A. Tuzhilin, in *Proceedings of the 2th ACM Conference on Recommender Systems*, RecSys 2008 (ACM, New York, NY, USA, 2008), pp. 11–18
32. T. Murakami, K. Mori, R. Orihara, in *New Frontiers in Artificial Intelligence*, ed. by K. Satoh, A. Inokuchi, K. Nagao, T. Kawamura, *Lecture Notes in Computer Science*, vol. 4914 (Springer, Berlin, Heidelberg, 2008), pp. 40–46
33. Y.C. Zhang, D.O. Séaghdha, D. Quercia, T. Jambor, in *Proceedings of the 5th ACM Conference on Web Search and Data Mining*, WSDM 2012 (ACM, New York, NY, USA, 2012), pp. 13–22
34. L. Chen, Y. Yang, N. Wang, K. Yang, Q. Yuan, in *Proceedings of the The World Wide Web Conference*, WWW 2019 (ACM, New York, NY, USA, 2019), pp. 240–250
35. D.M. Fleder, K. Hosanagar, Management Science **55**(5), 697–712 (2009)
36. Z. Szlávik, W. Kowalczyk, M. Schut, in *Proceedings of the 5th AAAI Conference on Weblogs and Social Media*, ICWSM 2011 (The AAAI Press, 2011)
37. D. Levinson, *Ethnic Groups Worldwide: A ready Reference Handbook* (Oryx Press, 1998)
38. G.P. Patil, C. Taillie, J. Am. Stat. Assoc. **77**(379), 548–561 (1982)
39. F. Van Dyke, in *Conservation Biology: Foundations, Concepts, Applications* (Springer Netherlands, Dordrecht, 2008), pp. 83–119

40. R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, in *Proceedings of the 2nd ACM Conference on Web Search and Data Mining*, WSDM 2009 (ACM, New York, NY, USA, 2009), pp. 5–14
41. J. Carbonell, J. Goldstein, in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 1998 (ACM, New York, NY, USA, 1998), pp. 335–336
42. O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, S.L. Wu, Information Retrieval **14**(6), 572–592 (2011)
43. H. Chen, D.R. Karger, in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2006 (ACM, New York, NY, USA, 2006), pp. 429–436
44. C.L. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, I. MacKinnon, in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2008 (ACM, New York, NY, USA, 2008), pp. 659–666
45. R.L. Santos, C. Macdonald, I. Ounis, in *Proceedings of the 19th International Conference on World Wide Web*, WWW 2010 (ACM, New York, NY, USA, 2010), pp. 881–890
46. C.X. Zhai, W.W. Cohen, J. Lafferty, in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2003 (ACM, New York, NY, USA, 2003), pp. 10–17
47. C.L. Clarke, N. Craswell, I. Soboroff, in *Proceedings of the 19th Text REtrieval Conference*, TREC 2010 (National Institute of Standards and Technology (NIST), 2010)
48. S. Vargas, P. Castells, D. Vallet, in *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2011 (ACM, New York, NY, USA, 2011), pp. 1211–1212
49. S. Vargas, P. Castells, D. Vallet, in *Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2012 (ACM, New York, NY, USA, 2012), pp. 75–84
50. G. Adomavicius, A. Tuzhilin, IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
51. M. Zhang, N. Hurley, in *Proceedings of the 2nd ACM Conference on Recommender Systems*, RecSys 2008 (ACM, New York, NY, USA, 2008), pp. 123–130
52. A. Veloso, M. Ribeiro, A. Lacerda, E. Moura, I. Hata, N. Ziviani, Special Issue on Novelty and Diversity in Recommender Systems, ACM Trans. Inf. Syst. Tech. **5**(4) (2014)
53. J.S. Breese, D. Heckerman, C. Kadie, in *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, UAI 1998 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998), pp. 43–52
54. M. Ge, C. Delgado-Battenfeld, D. Jannach, in *Proceedings of the 4th ACM Conference on Recommender systems*, RecSys 2010 (ACM, New York, NY, USA, 2010), pp. 257–260
55. J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 1999 (ACM, New York, NY, USA, 1999), pp. 230–237
56. M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, R. Burke, in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP 2020 (ACM, New York, NY, USA, 2020), pp. 154–162
57. D. Jannach, L. Lerche, G. Gedikli, G. Bonnin, in *Proceedings of the 21st International Conference on User Modeling, Adaptation and Personalization* (Springer, 2013), pp. 25–37
58. S. Vargas, P. Castells, in *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys 2014 (ACM, New York, NY, USA, 2014), pp. 145–152
59. J. Sanz-Cruzado, P. Castells, in *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys 2018 (Vancouver, Canada, 2018), pp. 233–241
60. Y. Zhang, J. Callan, T. Minka, in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2002 (ACM, New York, NY, USA, 2002), pp. 81–88
61. R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, 2nd edn. (Addison-Wesley Publishing Company, USA, 2008)

62. J. Allan, C. Wade, A. Bolivar, in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2003 (ACM, New York, NY, USA, 2003), pp. 314–321
63. S. Vargas, L. Baltrunas, A. Karatzoglou, P. Castells, in *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys 2014 (ACM, New York, NY, USA, 2014), pp. 209–216
64. H. Steck, in *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys 2018 (ACM, New York, NY, USA, 2018), pp. 154–162
65. T. Deselaers, T. Gass, P. Dreuw, H. Ney, in *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR 2009 (ACM, New York, NY, USA, 2009), pp. 39:1–39:8
66. M. Drosou, E. Pitoura, IEEE Data Eng. Bull. **32**(4), 49–56 (2009)
67. M. Drosou, K. Stefanidis, E. Pitoura, in *Proceedings of the 3rd ACM Conference on Distributed Event-Based Systems*, DEBS 2009 (ACM, New York, NY, USA, 2009), pp. 6:1–6:12
68. L. McGinty, B. Smyth, in *Proceedings of the 5th International Conference on Case-based Reasoning*, ICCBR 2003 (Springer, Berlin, Heidelberg, 2003), pp. 276–290
69. M. Drosou, E. Pitoura, Comparing diversity heuristics. Tech. rep., Technical Report 2009-05. Computer Science Department, University of Ioannina (2009)
70. K. Alodhaibi, A. Brodsky, G.A. Mihaila, in *Proceedings of the 43rd Hawaii International Conference on System Sciences*, HICSS 2010 (IEEE Computer Society, Washington, DC, USA, 2010), pp. 1–10
71. D. McSherry, in *Advances in Case-Based Reasoning*, ed. by S. Craw, A. Preece, (Springer, Berlin, Heidelberg, 2002), pp. 219–233
72. C. Yu, L. Lakshmanan, S. Amer-Yahia, in *Proceedings of the 12th International Conference on Extending Database Technology*, EDBT 2009 (ACM, New York, NY, USA, 2009), pp. 368–378
73. Q. Wu, F. Tang, L. Li, L. Barolli, I. You, Y. Luo, H. Li, in *Proceedings of the 26th IEEE Conference on Advanced Information Networking and Applications*, AINA 2012 (IEEE, 2012), pp. 191–198
74. J. Rao, A. Jia, Y. Feng, D. Zhao, in *Web Information Systems Engineering – WISE 2013*, ed. by X. Lin, Y. Manolopoulos, D. Srivastava, G. Huang, *Lecture Notes in Computer Science*, vol. 8180 (Springer, Berlin, Heidelberg, 2013), pp. 209–218
75. A. Bessa, A. Veloso, N. Ziviani, in *String Processing and Information Retrieval*, ed. by O. Kurland, M. Lewenstein, E. Porat, *Lecture Notes in Computer Science*, vol. 8214 (Springer International Publishing, 2013), pp. 17–28
76. M. Drosou, E. Pitoura, Proc. VLDB Endowment **6**(1), 13–24 (2012)
77. G. Adomavicius, Y. Kwon, in *Proceedings of the 1st ACM RecSys Workshop on Novelty and Diversity in Recommender Systems* (2011), DiveRS 2011, pp. 3–10
78. Y. Kwon, J. Intell. Inf. Syst. **18**(3), 119–135 (2012)
79. M. Zhang, N. Hurley, in *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT 2009 (IEEE Computer Society, Washington, DC, USA, 2009), pp. 508–515
80. R. Boim, T. Milo, S. Novgorodov, in *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, CIKM 2011 (ACM, New York, NY, USA, 2011), pp. 739–744
81. X. Li, T. Murata, in *Proceedings of the 7th International Conference on Computer Science & Education*, ICCSE 2012 (IEEE, 2012), pp. 905–910
82. L. Shi, in *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys 2013 (ACM, New York, NY, USA, 2013), pp. 57–64
83. J.B. Schafer, J.A. Konstan, J. Riedl, in *Proceedings of the 11th ACM Conference on Information and Knowledge Management*, CIKM 2002 (ACM, New York, NY, USA, 2002), pp. 43–51
84. M.T. Ribeiro, A. Lacerda, A. Veloso, N. Ziviani, in *Proceedings of the 6th ACM Conference on Recommender Systems*, RecSys 2012 (ACM, New York, NY, USA, 2012), pp. 19–26

85. N.J. Hurley, in *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys 2013 (ACM, New York, NY, USA, 2013), pp. 379–382
86. R. Su, L. Yin, K. Chen, Y. Yu, in *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys 2013 (ACM, New York, NY, USA, 2013), pp. 415–418
87. S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, UAI 2009 (AUAI Press, Arlington, VA, USA, 2009), pp. 452–461
88. D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac, User Model. User Adap. Inter. **25**, 427–491 (2015)
89. J. Wasilewski, N. Hurley, in *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP 2019 (ACM, New York, NY, USA, 2019), pp. 144–148. https://doi.org/10.1145/3320435.3320468
90. L. Iaquinta, M. de Gemmis, P. Lops, G. Semeraro, M. Filannino, P. Molino, in *Proceedings of the 8th Conference on Hybrid Intelligent Systems*, HIS 2008 (IEEE, 2008), pp. 168–173
91. K. Oku, F. Hattori, in *Proceedings of the 1st ACM RecSys Workshop on Novelty and Diversity in Recommender Systems*, DiveRS 2011 (2011)
92. U. Bhandari, K. Sugiyama, A. Datta, R. Jindal, in *Information Retrieval Technology*, ed. by R.E. Banchs, F. Silvestri, T.Y. Liu, M. Zhang, S. Gao, J. Lang, *Lecture Notes in Computer Science*, vol. 8281 (Springer, Berlin, Heidelberg, 2013), pp. 440–451
93. M. Taramigkou, E. Bothos, K. Christidis, D. Apostolou, G. Mentzas, in *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys 2013 (ACM, New York, NY, USA, 2013), pp. 335–338
94. K. Niemann, M. Wolpers, in *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD 2013 (ACM, New York, NY, USA, 2013), pp. 955–963
95. F. Mourão, C. Fonseca, C. Araújo, W. Meira, in *Proceedings of the 1st ACM RecSys Workshop on Novelty and Diversity in Recommender System*, DiveRS 2011 (2011)
96. M. Ge, D. Jannach, F. Gedikli, M. Hepp, in *Proceedings of the 14th International Conference on Enterprise Information Systems*, ICEIS 2012 (SciTePress, 2012), pp. 201–208
97. R. Hu, P. Pu, in *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI 2011 (ACM, New York, NY, USA, 2011), pp. 347–350
98. S. Castagnos, A. Brun, A. Boyer, in *Proceedings of the 3rd International Conference on Advances in Information Mining and Management*, IMMM 2013 (IARIA, Lisbon, Portugal, 2013), pp. 44–50
99. M.J. Welch, J. Cho, C. Olston, in *Proceedings of the 20th International Conference on World Wide Web*, WWW 2011 (ACM, New York, NY, USA, 2011), pp. 237–246
100. J. He, E. Meij, M. de Rijke, J. Assoc. Inf. Sci. Tech. **62**(3), 550–571 (2011)
101. J. Wang, in *Proceedings of the 31st European Conference on Information Retrieval*, ECIR 2009 (Springer, Berlin, Heidelberg, 2009), pp. 4–16
102. B. Carterette, in *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2011 (ACM, New York, NY, USA, 2011), pp. 903–912
103. A. Moffat, J. Zobel, ACM Trans. Inf. Syst. **27**(1), 2:1–2:27 (2008)
104. J. Wasilewski, N. Hurley, in *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys 2016 (ACM, New York, NY, USA, 2016), pp. 39–42
105. J. Wasilewski, N. Hurley, in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (ACM, New York, NY, USA, 2017), pp. 71–76
106. F. Radlinski, S. Dumais, in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2006 (ACM, New York, NY, USA, 2006), pp. 691–692
107. D. Vallet, P. Castells, in *Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2012 (ACM, New York, NY, USA, 2012), pp. 841–850

108. H. Steck, in *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys 2011 (ACM, New York, NY, USA, 2011), pp. 125–132
109. D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac, User Model. User Adap. Inter. **25**(5), 427–491 (2015)
110. A. Bellogín, P. Castells, I. Cantador, Information Retrieval **20**(6), 606–634 (2017)
111. R. Cañamares, P. Castells, in *Proceedings of the 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2018 (ACM, New York, NY, USA, 2018), pp. 415–424
112. A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E.H. Chi, C. Goodrow, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD 2019 (ACM, New York, NY, USA, 2019), pp. 2212–2220
113. W. Liu, J. Guo, N. Sonboli, R. Burke, S. Zhang, in *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys 2019 (ACM, New York, NY, USA, 2019), pp. 467–471
114. R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, F. Diaz, in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM 2018 (ACM, New York, NY, USA, 2018), pp. 2243–2251
115. N. Sonboli, F. Eskandanian, R. Burke, W. Liu, B. Mobasher, in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP 2020 (ACM, New York, NY, USA, 2020), pp. 239–247
116. F. Diaz, B. Mitra, M.D. Ekstrand, A.J. Biega, B. Carterette, in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM 2020 (ACM, New York, NY, USA, 2020), pp. 275–284
117. R. Baeza-Yates, Commun. ACM **61**(6), 54–61 (2018)
118. C. DiCiccio, S. Vasudevan, K. Basu, K. Kenthapadi, D. Agarwal, in *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD 2020 (ACM, New York, NY, USA, 2020), pp. 1467–1477
119. R. Epstein, R.E. Robertson, Proc. Natl. Acad. Sci. **112**(33), E4512–E4521 (2015)
120. A. Singh, T. Joachims, in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD 2018 (ACM, New York, NY, USA, 2018), pp. 2219–2228
121. P. Cremonesi, Y. Koren, R. Turrin, in *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys 2010 (ACM, New York, NY, USA, 2010), pp. 39–46
122. R. Cañamares, P. Castells, in *Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2017 (ACM, New York, NY, USA, 2017), pp. 215–224
123. A.J.B. Chaney, B.M. Stewart, B.E. Engelhardt, in *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys 2018 (ACM, New York, NY, USA, 2018), pp. 224–232
124. J.M. Hernández-Lobato, N. Houlsby, Z. Ghahramani, in *Proceedings of the 31st International Conference on Machine Learning*, ICML 2014 (Proc. of Machine Learning Research, Sheffield, UK, 2014), pp. 1512–1520
125. T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, T. Joachims, in *Proceedings of the 33rd International Conference on Machine Learning*, ICML 2016 (Proc. of Machine Learning Research, Sheffield, UK, 2016), pp. 1670–1679
126. D. Liu, P. Cheng, Z. Dong, X. He, W. Pan, Z. Ming, in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2020 (ACM, New York, NY, USA, 2020), pp. 831–840
127. A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, S. Dollé, in *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, WSDM 2018 (ACM, New York, NY, USA, 2018), pp. 198–206
128. A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, B. Carterette, in *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, WSDM 2019 (ACM, New York, NY, USA, 2019), pp. 420–428

129. A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudík, J. Langford, D. Jose, I. Zitouni, in *Proceedings of the 31st Conference on Neural Information Processing Systems*, NIPS 2017 (Curran Associates, Inc., Red Hook, NY, USA, 2017), pp. 3635–3645
130. L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie, D. Estrin, in *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys 2018 (ACM, New York, NY, USA, 2018), pp. 279–287
131. P.R. Rosenbaum, D.B. Rubin, Biometrika **70**(1), 41–55 (1983)
132. G.K. Patro, A. Biswas, N. Ganguly, K.P. Gummadi, A. Chakraborty, in *Proceedings of the Web Conference*, WWW 2020 (ACM/IW3C2, 2020), pp. 1194–1204
133. N. Helberger, K. Karppinen, L. D'Acunto, Inf. Commun. Soc. **21**(2), 191–207 (2018)
134. Y. Hu, Y. Koren, C. Volinsky, in *Proceedings of the 8th IEEE International Conference on Data Mining*, ICDM 2008 (IEEE Computer Society, Washington, DC, USA, 2008), pp. 263–272
135. A.P. Jeuland, in *Proceedings of the Educators' Conference*, 43 (American Marketing Association, 1978), pp. 33–37
136. F. Mourão, L. Rocha, C. Araújo, W. Meira, J. Konstan, Information Systems **71**, 137–151 (2017)
137. N. Tintarev, M. Dennis, J. Masthoff, in *User Modeling, Adaptation, and Personalization*, ed. by S. Carberry, S. Weibelzahl, A. Micarelli, G. Semeraro (Springer, Berlin, Heidelberg, 2013), pp. 190–202